Best Frame Selection in a Short Video

Jian Ren[1*] Xiaohui Shen[2*] Zhe Lin[3] Radomír Měch[3]
[1]Snap Inc. [2]ByteDance AI Lab [3]Adobe Research

Abstract

People usually take short videos to record meaningful moments in their lives. However, selecting the most representative frame, which not only has high image visual quality but also captures video content, from a short video to share or keep is a time-consuming process for one may need to manually go through all the frames in a video to make a decision. In this paper, we introduce the problem of the best frame selection in a short video and aim to solve it automatically. Towards this end, we collect and will release a diverse large-scale short video dataset that includes 11, 000 videos shoot in our daily life. All videos are assumed to be short (e.g., a few seconds) and each video has human-annotated of the best frame. Then we introduce a deep convolutional neural network (CNN) based approach with ranking objective to automatically pick the best frame from frame sequences extracted via short videos. Additionally, we propose new evaluation metrics, especially for the best frame selection. In experiments, we show our approach outperforms various other methods significantly.

1. Introduction

Nowadays, with benefits from the advances in hardware of digital cameras (e.g., the high-quality camera on iPhone) and an expanding number of social media platforms, more and more people would like to take short videos to record meaningful moments happened in various events, such as vacations, parties, and festivals. By using mobile devices, people can easily take high-quality videos and share them on social media via various applications. To represent a video, a thumbnail is usually selected and used to attract viewers by the first impression. The chosen thumbnail is considered as the best frame in a video, and it has an important role, for example, people care about thumbnails when browsing videos [2], so the best frame serves as a critical factor for determining whether to watch a video or not [8]. The best frame should have a high visual quality to be attractive and contain the essence of a video. Different from videos with a few minutes contain changing subjects, short videos, which may only a few seconds, always record one event. Therefore, one representative frame is usually enough to capture the content of a video while several frames are needed to show the content for video with a few minutes.

However, selecting the best frame from short videos to keep, share, or post is still a tedious, time-consuming, and challenging process. One needs to manually go through all frames one by one to make a decision. Additionally, many frames share similar content and visual image quality, which makes it even harder to decide the better one.

There have been studies on video summarization [54, 58, 32] where key shots including a few continuous frames are chosen to shorten videos into a compact version. However, for video summarization, frames with similar content in selected key shots are treated equally, and there is no ranking information among those frames to help decide the best one. Similarly, the works on image quality assessment [34, 33] give image scores on an absolute scale such as low quality or high quality, cannot still differentiate frames with similar quality, which is a common situation in short videos. Other works process image attributes learning as a regression problem [55, 27], but the context of a video is an essential factor for human preference on the best frame in short videos, which are ignored in those works. The learning of a robust ranking model for frame selection requires the understanding of low-level image features and high-level image features unified in one model. For example, some videos contain frames with similar content, low-level features such as undesired blur or out of focus are usually the decisive factors for determining best frames, while for other videos containing frames with changing context, the selection criteria could be high-level features such as image composition or aesthetics.

Given there is quite limited work on selecting one best frame from short videos, in this article, we aim to solve the problem of the best frame selection in a short video. To facilitate the study on best frame selection, we collect and will release a new dataset including 11, 000 diverse short videos that captured in our daily life. Each video has its frames extracted to form a frame sequence, and the best frame in a frame sequence is manually annotated. We present two

[*] Work done while at Adobe.

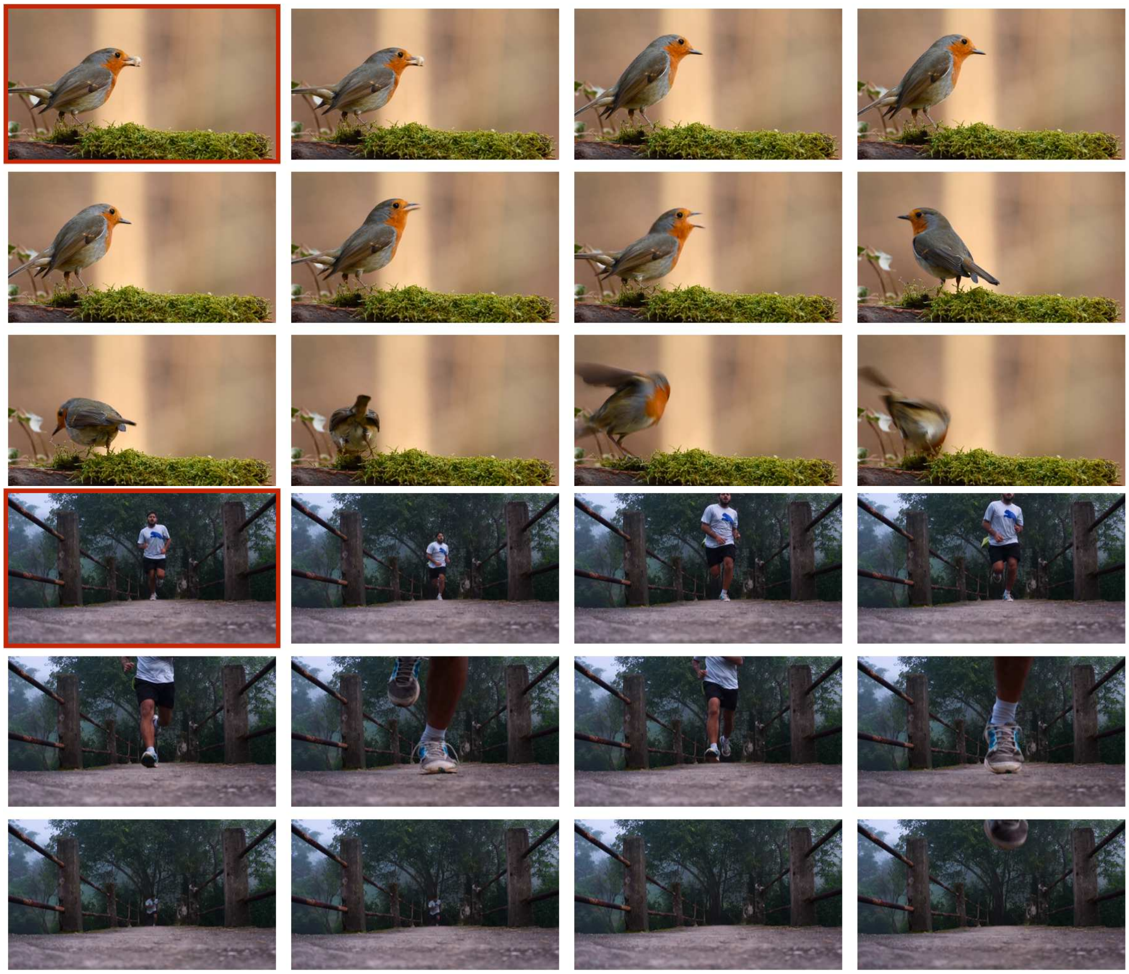


Figure 1: Two example short videos. The frames in each example are sampled from a short video and sorted from high to low in terms of their ground truth scores. In each frame sequence, the frame with red boundary is selected as the best frame by our approach. The selected frame in each short video captures the video's essence and has high visual quality.

short video examples from our dataset in Figure 1 where the frames are sorted from low to high based on their ground truth scores. The image with a red boundary is automatically selected as the best frame by our approach, which represents the content of the short videos and has high visual quality. We introduce an end-to-end training of a deep convolutional neural network (CNN) with a ranking loss function to choose the best frame from frame sequence automatically. Moreover, as the current study has revealed the importance of facial features in selecting images from albums [59], we incorporate facial features into one end-to-end network. During inference time, the network predicts frame scores for a given frame sequence, and the frame with the highest score is determined as the best frame. To evaluate the proposed method, we further introduce three evaluation metrics because exiting evaluation metrics for ranking problems, such as Spearman rank correlation [52], cannot determine whether the top frame is the best. Experimental results demonstrate that our approach has a clear advantage over baseline and prior works by outperforming them significantly.

In summary, our main contributions are three-fold:

- We propose and address the problem of best frame selection in a short video, and collect a large-scale short video dataset to facilitate research in this direction.
- We introduce learning strategies for studying the best frame selection in a short video with newly evaluation metrics.
- We incorporate facial features in the learning algorithm, which is effective in a real-world application on personal videos.

2. Related Work

2.1. Photo Triage

Photo selection has drawn more attention because of the increasing the of personal albums [7, 49]. Previous works have explored designing a user interface to help viewers interactively control the process of photo triage according to their preference [13, 23]. Recently, Chang et al. [4] proposed a neural network-based method to classify a pair of images in a photo collection to determine which one has higher quality. However, the method is limited in the following three aspects: (1) the model is trained on photo collections where each collection only has few images instead of videos, so it may not be optimal for best frame selection in short videos; (2) it has not incorporated face information in the neural network model; (3) the model is trained with a pair-wise classification loss which can be sensitive to pairs of training images with similar qualities, and it lacks efficiency during inference as it has to evaluate every pair.

Other studies on the assessment of image quality are alternative approaches for selecting photos from the collection. The estimated image score can be used to rank images in a collection that can be obtained from low-level image features, such as color [41], texture [9, 48] and lighting [37, 3, 24, 56], and high-level image features including aesthetics [11, 43, 33, 34, 27, 39], composition [37, 1, 17], content [36, 24], memorability [22, 25, 14], and interestingness [15, 11, 18]. However, the quality assessment across all the videos can fail in cases where videos contain images with similar attributes. Instead of using quality attributes, we find it is more effective to learn the ranking of frames.

2.2. Video Summarization and Keyframes Selection

Researchers have worked on video summarization with a primary goal of selecting representative key shots from videos. Recent studies show supervised video summarization [54, 58, 38, 35, 30, 42, 16, 57] that based on human-created summary sub-shots to learn selection criteria achieves better results than unsupervised summarization [26, 46, 53, 40, 32] which selects key shots according to manually designed criteria such as representativeness, diversity, and coverage. Similar to video summarization, the studies on keyframes selection aim to predict representative frames that are visually attractive and relevant to video content. Studies have explored the relationship between representative frames and various visual features [45, 12, 29, 44].

Different between the research on video summarization and keyframes selection, we work on short videos that are less than 10 seconds instead of videos of a few minutes. Furthermore, we aim to select the most representative frame from short videos instead of selecting a few key shots or frames. To achieve the goal, we collect a large short video dataset and propose evaluation metrics to assess methods targeting the problem.

3. Data Collection

Due to there is no public benchmark to solve the problem of best frame selection, we show how we build a stock video clip dataset (SVCD) primarily to address the issue. To the best of our knowledge, SVCD is the first video dataset that only focuses on short videos. We also compare SVCD with other datasets to show its advantages on the best frame selection. We will release SVCD and our models.

3.1. Short Videos Collection

We collect short videos by collecting videos from Adobe Stock Videos[1]. Each video has a manually annotated keywords list showing its content. To curate suitable short videos, we use a positive keyword list including 26 words, such as family, kid, and boy, to filter out videos captured in our daily life and a negative list including 136 words, such as white, background, and design, to block inappropriate ones. To obtain positive and negative keyword lists, we firstly sampled a small number of videos. If the video shows a moment that happens in our daily life, then we record the keywords of the video as positive keywords; otherwise, we record the keywords from it as negative keywords. We do this iteratively until we collect enough videos according to the keywords list. More specifically, the chosen videos in SVCD need to satisfy three requirements: 1) including at least one keyword from the positive keyword list; 2) not including any keyword from the negative keyword list; 3) the videos are no longer than 10 seconds.

In total, we collect 11, 000 short videos. Among each of them, we extract frames (e.g., 8 FPS) from it and uniformly sample 19 frames to form the frame sequence. For all videos, we use a face detection network [31] to detect face, which results in 5, 576 videos contain at least one frame with face. The total number of frames with face is 73, 170, which counts for 35.01% for all frames in SVCD. Additionally, we show the number of videos with the same number of frames that include face in Figure 2. 2, 422 videos have all 19 frames with detected face. For videos with face, the average number of frames with face is 13.1.

3.2. Datasets Annotation

To get human preference on the most representative frames in short videos, we present the sampled frame sequence in each video to Amazon Mechanical Turk[2] (AMT) to collect annotations for each frame. For each task on AMT, only workers who pass the qualification test can work on our tasks to label 19 frames from each video with a score

[1] https://stock.adobe.com/video

[2] www.mturk.com

Table 1: Datasets comparison of SVCD with benchmarks on video summarization, keyframes selection, and photo triage.

Dataset	number of videos (series)	video time	average number of photos
SumeMe [19]	25	from 1.5 to 6.5 minutes	-
TVSum [46]	50	from 1 to 5 minutes	-
OVP [10]	50	from 1 to 4 minutes	-
Youtube [10]	50	from 1 to 10 minutes	-
Yahoo Screen [45]	1, 118	average 2.8 minutes	-
Photo triage [4]	5, 953	-	2.6
SVCD	11, 000	less than 10 seconds	19

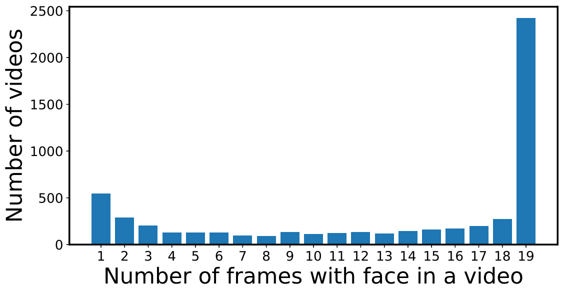


Figure 2: The number of videos that have the same number of frames with face.

in $[1, 2, 3]$, where 3 indicates a frame can be used to represent the video and has high visual image quality while 1 denotes a frame with low visual image quality and does not capture video's content. To control annotation quality, we randomly insert short videos with known ground truth into the AMT tasks. Those videos have obviously best frame so attentive workers should find out the best frame. Each video is labeled by five distinct AMT workers, and we use the weighted averaged ratings over the five AMT workers as ground truth score, in which the weights are proportional to workers' accuracy on the quality control videos and normalized among five workers. In total, SVCD consists of 209, 000 labeled images.

3.3. Datasets Comparison

To show the advantages of SVCD, we compare it with other benchmark datasets from video summarization, which includes SumMe [19], TVSum [46], Open Video Project (OVP) [10], and Youtube [10], a Yahoo Screen dataset from keyframes selection [45], and a triage dataset from photo triage [4]. We show the number of videos or photo series contained in each dataset and the averaged number of photos in each video in Table 1. Compared with the datasets from video summarization and keyframes selection, SVCD includes more videos. Additionally, the averaged number of photos in each frame series is 7 times bigger compared with the photo triage dataset [4]. Such comparison shows SVCD is superior to other datasets and can be accepted as a benchmark for studies of the best frame selection in short videos for it includes a large number of videos, and all of them are short (e.g., less than 10 seconds).

4. Methods

In this section, we propose the model to estimate the frame score in short videos by considering both the representatives and facial features of frames. During the inference time, the best frame of a short video is selected by choosing the one with the highest score. Given SVCD contains videos shot in our daily life under various contents and quality, learning to estimate the accurate rank of frames in each video with a ranking objective is easier and more favorable than learning a regression by using the Euclidean loss to force the frame score to get close to ground-truth annotation. Therefore, we adopt the idea of Siamese network [6] where pairs of images are given as input during the training process and optimize the network by using a ranking loss function. Furthermore, SVCD concentrates on the videos that captured in our daily lives. For those kinds of videos, the face is an import factor when people consider selecting the best frame. Inspired by an existing study that facial features play an essential role in selecting representative photos from collection [59], we incorporate prior knowledge obtained from faces in one end-to-end network for determining frame scores. The architecture is shown in Figure 3, and the details are introduced as follows.

4.1. Siamese CNN

When selecting the best frame from a video, people follow the criterion that the selected frame can represent the video content, and it has high image quality. However, the various quality of videos may bias the annotation for AMT workers. Thus rather than learning a regression to predict

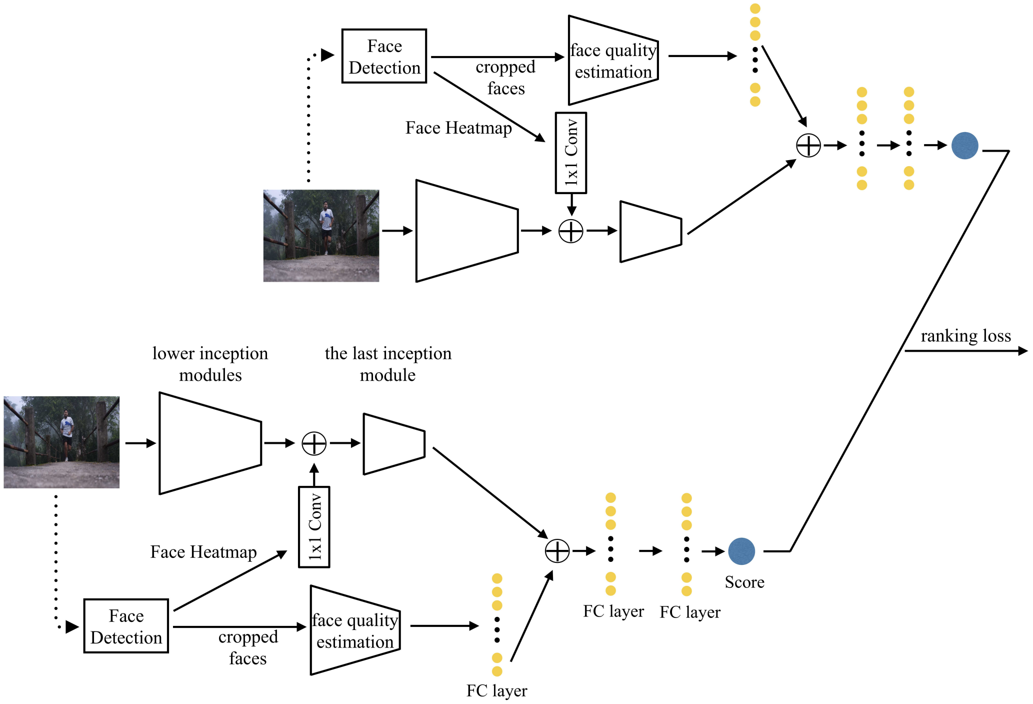


Figure 3: The proposed network for predicting best frames from short videos. The siamese network uses pairs of frames sampled from the same video as input. Facial features, including face heatmaps and face quality, are incorporated into the network.

an absolute score for frames in all videos, which is difficult, learning the ranking of frames within each video is more suitable in applications especially short videos may contain frames with similar context and quality. Toward this end, we utilize a Siamese CNN architecture that includes two pathways with shared parameters where each pathway follows GoogLeNet [47] architecture. To learn the parameters, we adopt the loss function as Piecewise Ranking (PR) loss [50] because it introduces relaxation of ground truth score and makes the network more stable for a subjective task. Supposing input pair of frames fed into the network is $(\mathrm{I}_1, \mathrm{I}_2)$, $G(\mathrm{I}_i)$ is ground truth score of the frame I_i, and $P(\mathrm{I}_i)$ is its estimated value from the network. The PR loss is formed as the following:

$$\mathrm{PR} = \begin{cases} \frac{1}{2}\max(0, |D_\mathrm{p}| - \mathrm{m_s})^2 & \text{if } D_\mathrm{g} < \mathrm{m_s} \\ \frac{1}{2}\left\{\max(0, \mathrm{m_s} - D_\mathrm{p})^2 + \max(0, D_\mathrm{p} - \mathrm{m_d})^2\right\} & \\ & \text{if } \mathrm{m_s} \le D_\mathrm{g} \le \mathrm{m_d} \\ \frac{1}{2}\max(0, \mathrm{m_d} - D_\mathrm{p})^2 & \text{if } D_\mathrm{g} > \mathrm{m_d} \end{cases}, \tag{1}$$

where $D_\mathrm{g} = G(\mathrm{I}_1) - G(\mathrm{I}_2)$ is the ground truth score difference between the two input frames, and $D_\mathrm{p} = P(\mathrm{I}_1) - P(\mathrm{I}_2)$ is the predicted value difference. $\mathrm{m_s}$ and $\mathrm{m_d}$ are constant margin values. We denote the network design as *Siamese CNN*. The *Siamese CNN* is fine-tuned from an off-the-shelf image classification or tagging CNN model. For the inference, *Siamese CNN* estimates frame scores for a short video and the higher score denotes the frame is more representative in the frame sequence.

4.2. Incorporating Facial Features

Many videos that are taken in our daily events contain people. For those videos that include human faces, the size, location, and quality of faces are important crucial when determining the representativeness of frames. For example, people would like to choose the frame that has an intact face instead of the one with the only partial face, and the frame contains face with high visual quality is more appealing and attractive than the one with low image quality (e.g., face blurriness). Therefore, we incorporate face information into *Siamese CNN* by using face heatmap to represent the size and location of the face and face quality feature to imply face quality.

Face Heatmap CNN. To generate the face heatmap, we first use a state-of-the-art face detection network [31] to detect faces in frames. Frames without rescaling are forwarded into the face detection network. Then we use the

Gaussian kernel to represent the size and location of the face, which are inferred from the coordinates of face boundary, on face heatmap. For frames without a human face, there is no face shown on face heatmap. The generated face heatmap has the same image size (e.g., height and width) as the frames forwarded into the face detection network. We show the input pair of frames and their corresponding face heatmaps in Figure 3.

To incorporate the knowledge from face heatmap into *Siamese CNN*, the input face heatmaps are passed through a convolutional layer with kernel size as 1×1 and number of channels as 384 before concatenated with the features maps obtained from lower inception modules. We denote the network design as *Face Heatmap CNN*. During the training process, we only train the last inception module and fully connected layers of *Face Heatmap CNN* from scratch using PR loss, while weights of lower inception modules of *Face Heatmap CNN* are fixed and the same as the weights in *Siamese CNN*.

Face Quality CNN. In order to get face quality features, we first train a face quality model to estimate face quality. We manually annotate the quality of faces in a face recognition dataset [20] and use score selected from $[0, 0.5, 1]$ to show face quality where the higher score implies the better face quality. Then we utilize SqueezeNet [21] as the network to learn face quality. The loss function is adopted as Euclidean loss (EL):

$$\text{EL} = \frac{1}{2}\sum_{i=1}^{N} \left\| g_i^2 - p_i^2 \right\|_2^2, \qquad (2)$$

where g_i is the ground truth annotation of face quality and p_i is the prediction score.

To incorporate face quality estimation to *Face Heatmap CNN*, the input pair of frames are forwarded to a face detection network [31] to get cropped faces. We then use the pre-trained SqueezeNet to get the neural activations from the second to the last layer as the face quality features. The face quality features are forwarded to a fully connected layer with 256 hidden units before concatenated with the feature maps from the last inception module. For frames with more the one faces, the averaged feature vectors are used; for frames without a face, feature vector with all zeros is used. We denote the method as *Face Quality CNN*. To train *Face Quality CNN*, we fix the lower inception modules where the weights are the same as *Siamese CNN* and initialize the last inception module from *Face Heatmap CNN*. The fully connected layers are trained from scratch.

5. Experiments and Results

In this section, we evaluate the proposed approach on SVCD and compare it with several other methods. We show qualitative examples in supplemental materials.

5.1. Implementation Details

Datasets preparation. For training and testing, we split SVCD by randomly select 1,000 videos as a testing set and others as a training set. Since videos comprise the various number of best frames, we exclude videos from the testing set that have all 19 frames with the same score and get 812 videos left in the testing set. The number of videos in the testing set that have the same number of best frames is shown in Figure 4. The average number of best frames for videos in the testing set is 7.46.

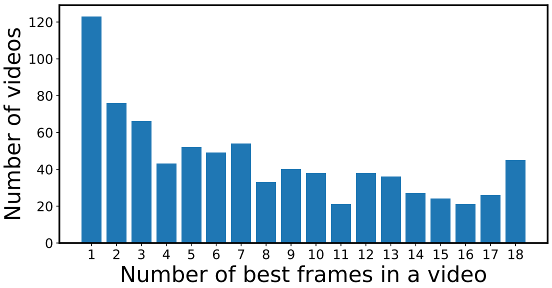


Figure 4: The number of videos in the testing set that have the same number of best frames.

Training details. During the training process, the ground truth scores of frames are normalized in 0 to 1. We set m_d = 0.03 and m_d = 0.1 in Equation 1 for PR loss. Since our training strategy includes three steps, we firstly train *Siamese CNN*. Following the data augmentation from [28], we resize input pair of images as 256×256, randomly crop them to 224×224, and apply horizontal flips. Then, to train *Face Heatmap CNN*, we fine-tune it from *Siamese CNN* where parameters in the lower inception modules are fixed. The input pair of images are resized as 224×224 without cropping, and their corresponding face heatmaps are resized as 7×7 before forwarded into the network. Lastly, we learn *Face Quality CNN* by fine-tuning it from *Face Heatmap CNN* and keep parameters in the lower inception modules fixed. To train the face quality model incorporated in *Face Quality CNN*, we resize input faces as 128×128 and randomly cropped them to 112×112. We train all three networks using mini-batch Stochastic Gradient Descent with batch size as 32, weight decay as 0.0002, and momentum as 0.9. The initial learning rate is 0.001 and multiplied with 0.96 after every 16,000 iterations.

5.2. Evaluation Metrics

To evaluate different approaches for selecting the best frame from short videos, we further propose new evaluation metrics. Although Spearman rank correlation [52] is widely utilized for assessing ranking models, which is calculated as the ranking correlation between predicted scores and ground truth annotations, we do not adopt it because for evaluating the selection of best frame, it is more important to tell whether the top frame given by prediction model is the best one or not rather than estimating the ranking for all frames in a video. Instead, we can calculate the differences between the selected frame and the best one in terms of their ground truth scores or the ranking and utilize those information for evaluation. Thus, we introduce new metrics.

The first evaluation metric is **Score Difference (SD)**. Supposing the k^{th} video includes n frames, and the ground truth score for frames are $S_{k_i}, ..., S_{k_n}$. Assuming the frame selected by the prediction model is S_{k_p}, and the best frame has score as S_{k_m}. Then score difference is defined as:

$$\text{SD} = \frac{1}{N}\sum_{k=1}^{N}(S_{k_m} - S_{k_p}), \tag{3}$$

where N is the total number of testing videos.

The second metric is **Rank Difference (RD)**. Rank difference is similar to score difference. When we sort the all frames in the k^{th} video from low to high in terms of their ground truth scores, we get a rank for each frame as $R_{k_i}, ..., R_{k_n}$. Assuming the frame selected by the prediction model has the rank as R_{k_p}, and the most representative frame has the rank as R_{k_m}. Then the rank difference is:

$$\text{RD} = \frac{1}{N}\sum_{k=1}^{N}(R_{k_m} - R_{k_p}). \tag{4}$$

The last one is **Accurate Percentage (AP)**. Let N_p be the number of videos that the frame selected by the prediction model is the best. The accurate percentage is defined as AP = $\frac{N_p}{N}$. For SD and RD, the model with lower values indicates better performance, while for AP, the model with a higher value has better performance.

5.3. Comparison Results

Based on three evaluation metrics, we compare our approach with the following methods.

Random selection: The method randomly selects a frame from a frame sequence. We run the random selection for ten times and report the average performance.

Middle selection: We use the method to pick a middle frame from a video when frames are sorted by their time stamp. The previous study has shown that middle selection performs better than selecting the first frame [12]. In experiments, we select the 10^{th} frame in short videos.

Image Aesthetics [27, 43]: In each video sequence, we try two aesthetics models to predict frame aesthetics and use the aesthetics score as frame score. The best frame is selected with the highest aesthetic value.

Photo Composition [5, 51]: We compare two photo composition studies, which are View Evaluation Net (VEN) [51] and View Finding Network (VFN) [5], for best frame selection. The predicted score from the two networks are used to decide the best frame.

Photo Triage [4]: We use the classification loss introduced [4] to train a Siamese network where a two-way Softmax is used to determine the better frame from an input pair of images. For a fair comparison, we use the same network architecture as *Siamese CNN*. The best frame in short video is determined through majority voting. The method achieves state-of-the-art results in photo triage [4], which is a challenging problem.

Euclidean Loss: We use a Euclidean loss function instead of Equation 1 to train a network.

The comparison results demonstrated in Table 2 show the proposed *Siamese CNN* has better performance than other works. For short videos, many frames have similar aesthetics quality, and content. So the aesthetics model [27, 43] cannot tell the differences between frames with similar aesthetics quality, which is consistent with findings from [5, 51] that the performance of aesthetics models on ranking images with similar views are not guaranteed. Also, those studies on learning photo composition [5, 51] performs better than the work on image aesthetics. Similarly, Euclidean loss do not perform well on the task of ranking views from similar images.

For the study on photo triage, the work [4] uses a Siamese network to perform the classification of input frames. However, a ranking loss is more favorable than a binary classification for a subjective task because it can in-

Table 2: Comparison of proposed methods with other approaches on SVCD. *Siamese CNN* achieves better performance than other methods and *Face Quality CNN* shows advantage over *Siamese CNN*.

Approach	SD	RD	AP
Random Selection	0.3800	5.8894	03930
Middle Selection [12]	0.3270	5.4938	0.4236
AesRankNet [27]	0.3862	5. 8445	0.3793
Image Aesthetics [43]	0.3633	5.4778	0.4027
VFN [5]	0.3070	4.9089	0.4495
VEN [51]	0.2152	3.9015	0.5234
Photo Triage [4]	0.2240	3.7418	0.5428
Euclidean Loss	0.2597	4.5234	0.4741
Siamese CNN	0.1832	3.1404	0.5874
Face Heatmap CNN	0.1771	3.0961	0.5961
Face Quality CNN	**0.1646**	**2.9581**	**0.6096**

troduce the relaxation of ground truth score. Furthermore, we incorporate face information in our network, which is not considered in photo triage [4].

Moreover, we show the ablation analysis of three proposed networks at the bottom of Table 2. *Face Quality CNN* achieves better results compared with *Face Heatmap CNN* and *Siamese CNN*, which indicates facial features including size, location, and quality of face are both important factors when considering the best frames in short videos. In Figure 5, we show more details of the network performance on videos with a different number of best frames. Videos with more best frames are usually more predictable than videos with less best frames.

Table 3: Comparison results between two sampling strategies. Best pairs sampling is beneficial for best frame selection than all pairs sampling.

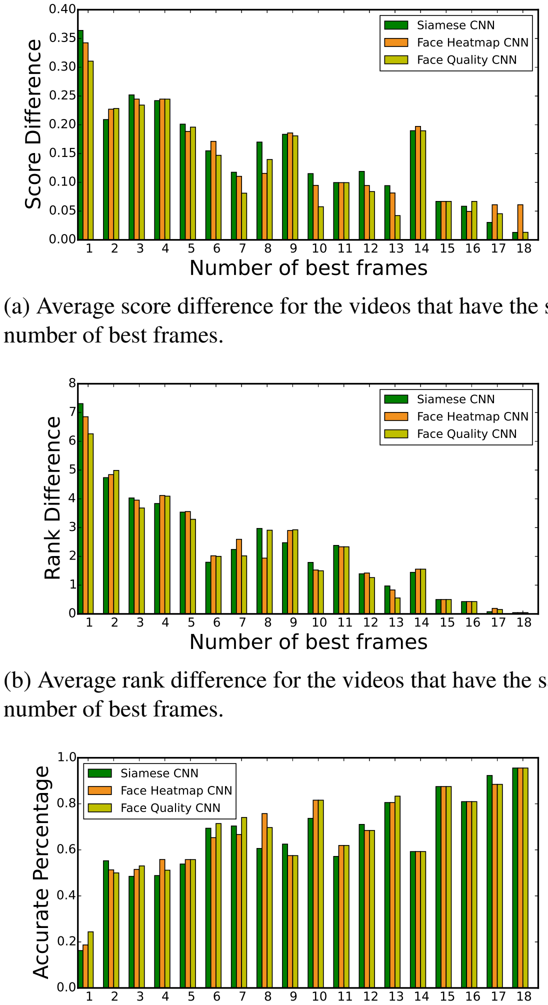
Sampling Strategy	SD	RD	AP
All pairs sampling	0.2239	3.7418	0.5428
Best pairs sampling	0.1832	3.1404	0.5874

5.4. Analysis on Sampling Strategies

We further analyze whether it is necessary to sample more pairs of frames to train the networks. We compare with two sampling strategies for learning *Siamese CNN*. The first one is **all pairs sampling** where for each video in the training set, we sample two frames as a pair when the two frames have different scores. The second one is **best pairs sampling** where for each pair of frames, it includes one frame is the best frame, and another is not. We compare the two sampling strategies on *Siamese CNN*. The results shown in Table 3 show the advantage of best pairs sampling as it has better performance than all pairs sampling on all three evaluation metrics. The comparison indicates it is more important to show networks the difference between the best frame and other frames rather than passing all pairs of frames into networks when train networks to learn the most representative frame from a video sequence.

6. Conclusion and Discussion

In this article, we introduce a challenging problem, which is the best frame selection in short videos. To facilitate the study, we collect a large dataset that includes 11, 000 short videos. Based on the dataset, we introduce an end-to-end learning model with a ranking objective to select the best frames by considering both the representativeness and facial features in frames. The proposed method outperforms existing studies significantly on the introduced evaluation metrics. However, for the short videos with dramatically changing content such as sports videos that record basketball games where some frames focus on one player and others focus on a team of players, the proposed method may not select the best frame that seizes video content. We have tried using a recurrent neural network to consider all frames in a short video when deciding the best frame. Unfortunately, we did not observe significant improvement. For future work, it would be interesting to investigate other factors that decide the selection of best frames.



(a) Average score difference for the videos that have the same number of best frames.

(b) Average rank difference for the videos that have the same number of best frames.

(c) Average accurate percentage for the videos that have the same number of best frames.

Figure 5: The score difference, rank difference, and accurate percentage for videos in the testing set with the different number of best frames.

References

[1] S. Bhattacharya, R. Sukthankar, and M. Shah. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM inter-*

national conference on Multimedia, pages 271–280. ACM, 2010.

[2] G. Buscher, E. Cutrell, and M. R. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 21–30. ACM, 2009.

[3] V. Bychkovsky, S. Paris, E. Chan, and F. Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *CVPR*, pages 97–104. IEEE, 2011.

[4] H. Chang, F. Yu, J. Wang, D. Ashley, and A. Finkelstein. Automatic triage for a photo series. *ACM Trans. Graph.*, 35(4):148:1–148:10, 2016.

[5] Y.-L. Chen, J. Klopp, M. Sun, S.-Y. Chien, and K.-L. Ma. Learning to compose with professional photographs on the web. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 37–45. ACM, 2017.

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005.

[7] W.-T. Chu and C.-H. Lin. Automatic selection of representative photo and smart thumbnailing using near-duplicate detection. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 829–832. ACM, 2008.

[8] S. J. Cunningham and D. M. Nichols. How people find videos. In *JCDL'08*, pages 201–210. ACM, 2008.

[9] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, pages 288–301. Springer, 2006.

[10] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, 32(1):56–68, 2011.

[11] S. Dhar, V. Ordonez, and T. L. Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR*, pages 1657–1664. IEEE, 2011.

[12] F. Dirfaux. Key frame selection to represent a video. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 2, pages 275–278. IEEE, 2000.

[13] S. Drucker, C. Wong, A. Roseway, S. Glenner, and S. De Mar. Photo-triage: Rapidly annotating your digital photographs. 2003.

[14] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem. What makes an object memorable? In *ICCV*, pages 1089–1097, 2015.

[15] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao. Interestingness prediction by robust learning to rank. In *ECCV*, pages 488–503. Springer, 2014.

[16] B. Gong, W.-L. Chao, K. Grauman, and F. Sha. Diverse sequential subset selection for supervised video summarization. In *NIPS*, pages 2069–2077, 2014.

[17] Y. Guo, M. Liu, T. Gu, and W. Wang. Improving photo composition elegantly: Considering image similarity during composition optimization. In *Computer graphics forum*, volume 31, pages 2193–2202. Wiley Online Library, 2012.

[18] M. Gygli, H. Grabner, H. Riemenschneider, F. Nater, and L. Van Gool. The interestingness of images. In *ICCV*, pages 1633–1640, 2013.

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

[20] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[21] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer. Squeezenet: Alexnet-level accuracy with 50× fewer parameters and$<$0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[22] P. Isola, D. Parikh, A. Torralba, and A. Oliva. Understanding the intrinsic memorability of images. In *NIPS*, pages 2429–2437, 2011.

[23] D. E. Jacobs, D. B. Goldman, and E. Shechtman. Cosaliency: Where people look when comparing images. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 219–228. ACM, 2010.

[24] L. Kaufman, D. Lischinski, and M. Werman. Content-aware automatic photo enhancement. In *Computer Graphics Forum*, volume 31, pages 2528–2540. Wiley Online Library, 2012.

[25] A. Khosla, A. Das Sarma, and R. Hamid. What makes an image popular? In *WWW*, pages 867–876. ACM, 2014.

[26] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *CVPR*, pages 2698–2705, 2013.

[27] S. Kong, X. Shen, Z. Lin, R. Mech, and C. Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, pages 662–679. Springer, 2016.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[29] W.-S. Lai, Y. Huang, N. Joshi, C. Buehler, M.-H. Yang, and S. B. Kang. Semantic-driven generation of hyperlapse from 360 degree video. *IEEE transactions on visualization and computer graphics*, 24(9):2610–2621, 2017.

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

[31] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015.

[32] T. Liu and J. R. Kender. Optimization algorithms for the selection of key frame sequences of variable length. In *ECCV*, pages 403–417. Springer, 2002.

[33] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang. Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 457–466. ACM, 2014.

[34] X. Lu, Z. Lin, X. Shen, R. Mech, and J. Z. Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *ICCV*, pages 990–998, 2015.

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

[36] W. Luo, X. Wang, and X. Tang. Content-based photo quality assessment. In *ICCV*, pages 2206–2213. IEEE, 2011.

[37] Y. Luo and X. Tang. Photo and video quality evaluation: Focusing on the subject. In *ECCV*, pages 386–399. Springer, 2008.

[38] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *CVPR*, volume 1, 2017.

[39] N. Murray, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *CVPR*, pages 2408–2415. IEEE, 2012.

[40] C.-W. Ngo, Y.-F. Ma, and H.-J. Zhang. Automatic video summarization by graph modeling. In *CVPR*, pages 104–109. IEEE, 2003.

[41] M. Nishiyama, T. Okabe, I. Sato, and Y. Sato. Aesthetic quality classification of photographs based on color harmony. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 33–40. IEEE, 2011.

[42] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid. Category-specific video summarization. In *ECCV*, pages 540–555. Springer, 2014.

[43] J. Ren, X. Shen, Z. L. Lin, R. Mech, and D. J. Foran. Personalized image aesthetics. In *ICCV*, pages 638–647, 2017.

[44] M. Silva, W. Ramos, J. Ferreira, F. Chamone, M. Campos, and E. R. Nascimento. A weighted sparse sampling and smoothing frame transition approach for semantic fast-forward first-person videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2383–2392, 2018.

[45] Y. Song, M. Redi, J. Vallmitjana, and A. Jaimes. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 659–668. ACM, 2016.

[46] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes. Tvsum: Summarizing web videos using titles. In *CVPR*, pages 5179–5187, 2015.

[47] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[48] H. Tang, N. Joshi, and A. Kapoor. Learning a blind measure of perceptual image quality. In *CVPR*, pages 305–312. IEEE, 2011.

[49] T. C. Walber, A. Scherp, and S. Staab. Smart photo selection: Interpret gaze as personal interest. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 2065–2074. ACM, 2014.

[50] Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller, and G. W. Cottrell. Event-specific image importance. In *CVPR*, pages 4810–4819, 2016.

[51] Z. Wei, J. Zhang, X. Shen, Z. Lin, R. Mech, M. Hoai, and D. Samaras. Good view hunting: Learning photo composition from dense view pairs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2018.

[52] A. D. Well and J. L. Myers. *Research design & statistical analysis*. Psychology Press, 2003.

[53] H. Yang, B. Wang, S. Lin, D. Wipf, M. Guo, and B. Guo. Unsupervised extraction of video highlights via robust recurrent auto-encoders. In *ICCV*, pages 4633–4641, 2015.

[54] T. Yao, T. Mei, and Y. Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *CVPR*, pages 982–990, 2016.

[55] N. Yu, X. Shen, Z. Lin, R. Mech, and C. Barnes. Learning to detect multiple photographic defects. *arXiv preprint arXiv:1612.01635*, 2016.

[56] L. Yuan and J. Sun. Automatic exposure correction of consumer photographs. In *ECCV*, pages 771–785. Springer, 2012.

[57] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*, pages 1059–1067, 2016.

[58] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *ECCV*, pages 766–782. Springer, 2016.

[59] J.-Y. Zhu, A. Agarwala, A. A. Efros, E. Shechtman, and J. Wang. Mirror mirror: Crowdsourcing better portraits. *ACM Transactions on Graphics (TOG)*, 33(6):234, 2014.