Fast Video Shot Transition Localization with Deep Structured Models

Shitao Tang(✉), Litong Feng, Zhanghui Kuang, Yimin Chen, and Wei Zhang

Sensetime Research, Hong Kong, China
shitaot@gmail.com

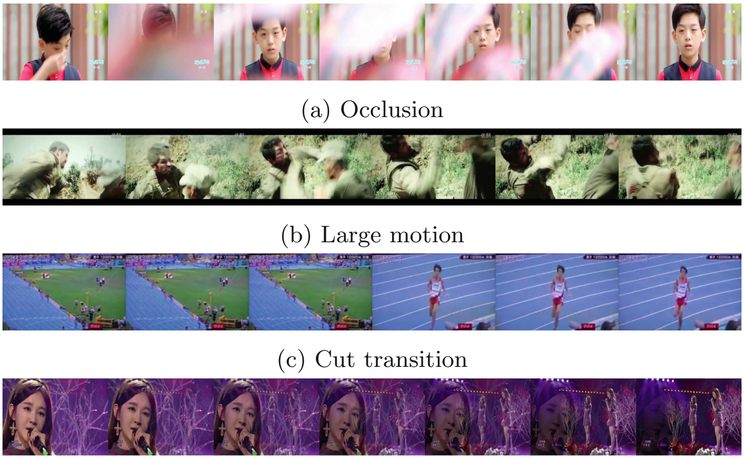
Abstract. Detection of video shot transition is a crucial pre-processing step in video analysis. Previous studies are restricted on detecting sudden content changes between frames through similarity measurement and multi-scale operations are widely utilized to deal with transitions of various lengths. However, localization of gradual transitions are still under-explored due to the high visual similarity between adjacent frames. Cut shot transitions are abrupt semantic breaks while gradual shot transitions contain low-level spatial-temporal patterns caused by video effects, e.g. dissolve. In this paper, we propose a structured network aiming to detect these two shot transitions using targeted models separately. Considering speed performance trade-offs, we design the following framework. In the first stage, a light filtering module is utilized for collecting candidate transitions on multiple scales. Then, cut transitions and gradual transitions are selected from those candidates by separate detectors. To be more specific, the cut transition detector focus on measuring image similarity and the gradual transition detector is able to capture temporal pattern of consecutive frames, even locating the positions of gradual transitions. The light filtering module can rapidly exclude most of the video frames from further processing and maintain an almost perfect recall of both cut and gradual transitions. The targeted models in the second stage further process the candidates obtained in the first stage to achieve a high precision. With one TITAN GPU, the proposed method can achieve a 30× real-time speed. Experiments on public TRECVID07 and RAI databases show that our method outperforms the state-of-the-art methods. To train a high-performance shot transition detector, we contribute a new database ClipShots, which contains 128636 cut transitions and 38120 gradual transitions from 4039 online videos. ClipShots intentionally collect short videos for more hard cases caused by hand-held camera vibrations, large object motions, and occlusion. The database is avaliable at https://github.com/Tangshitao/ClipShots.

Keywords: Shot boundary detection · Deep structured model · ClipShots

1 Introduction

Shot transition detector is a necessary component in many video recognition tasks [10,23,29]. The goal of shot transition detection is to find semantic breaks

C. V. Jawahar et al. (Eds.): ACCV 2018, LNCS 11361, pp. 577–592, 2019.
https://doi.org/10.1007/978-3-030-20887-5_36



(a) Occlusion

(b) Large motion

(c) Cut transition

(d) Gradual transition

Fig. 1. Challenge of shot boundary detection

in videos. Cut transitions are defined as abrupt transitions from one sequence to another while gradual transitions are almost the same but in a gradual manner. They share one common attribute, the start of a transition and the end of a transition are semantically different. Previous methods focus on finding both cut transitions and gradual transitions with one similarity function [25,27]. Such methods have shown a great success in cut transition detection in the aspects of both speed and accuracy. However, when applied to gradual transition detection, it is not effective in the detection of gradual transitions. As Fig. 1 shows, it is widely recognized that many large motions or occlusion, e.g. camera movement, are detected as positive when only measuring similarity. In order to overcome this shortcoming, recent research [9,17] begins to explore the temporal pattern of gradual transitions. Therefore, in [9], the C3D ConvNet is adopted to classify segments into three classes (cut, gradual and background), which achieves state-of-the-art performance. Yet C3D ConvNet not only consumes too much computing resources, but is also not an effective architecture for handling both cut and gradual transitions, i.e. the lengths of gradual transitions are varying but C3D ConvNet is not designed for multi-scale detection. Inspired by this method and previous similarity measurement method, we present a cascade framework, consisting of a targeted cut transition detector and a targeted gradual transition detector. The cut transition detector, for measuring the image similarity, is fast and accurate while the gradual transition detector is capable of capturing the temporal pattern of gradual transitions in multi-scale level. In addition, compared to deepSBD, our framework can locate both cut transitions and gradual transitions accurately.

In this work, we present a new cascade framework, a fast and accurate approach for shot boundary detection. The first stage applies a ridiculously fast method to initially filter the whole video and selects the candidate segments. This stage is for accelerating the framework (up to 2 times faster than not) and facilitate the training for the cut/gradual detector. In the second stage, we use a well designed 2D ConvNet learning the similarity function between two images to locate the cut transitions. The third stage utilizes a novel C3D ConvNet model to locate positions of gradual transitions. Typically, we use the notation of default boxes introduced in [15] and propose a novel single shot boundary detector (SSBD).

In sum, our framework is fast and accurate for shot boundary detection and achieves state-of-the-art performance on many public databases running at 700 FPS without any bells and whistles.

Current datasets, i.e. TRECVID and RAI, are not sufficient for training deep neural net due to limited dataset size. Besides, the training set is various in different work when evaluating supervised methods on TRECVID and RAI databases. For training a high performance neural network and a fair comparison between different methods, we contribute a new large-scale video shot database ClipShots consisting of different types of videos collected from Youtube and Weibo. ClipShots is the first large-scale database for shot boundary detection and will be released.

Aspects of novelty of our work include:

- We separate cut transition detection and gradual transition detection, designing targeted network structures with different purposes.
- We design a cascade framework for accelerating the processing speed.
- We collect the first large-scale database for shot boundary detection training and evaluation.

2 Related Work

In this section, we introduce the work related to our proposed framework.

Unsupervised Shot Boundary Detection Method. In decades, many researchers explore to design similarity function finding transitions with hand-crafted features. In [27], average Intensity Measurement (AIM), Histogram Comparison (HC), Likelihood Ratio (LR) is used as the feature extractor. It is observed that similarities often vary gradually within a shot but abruptly in shot boundaries so the paper proposes an adaptive threshold should be applied when selecting positive samples. This method greatly improves the gradual transition performance compared to methods that only use static thresholds. Besides, another benefit is that it runs very fast so we integrate it in our framework to select potential shot boundaries. Yuan et al. [25] proposes a graph partition model to perform temporal data segmentation. It treats every frame as a node and calculate the similarity metrix and the scores of the cuts, selecting feasible cuts whose scores are the local minima of the corresponding neighborhoods. These two methods all rely on well designed hand-crafted features to calculate the similarity of two images.

Supervised Shot Boundary Detection Method. Due to the shortcoming of unsupervised methods, Yuan et al. [26] adopts a supervised way, a support vector machine trained to classify different shot boundaries with extracted features. In [16], shot boundaries are classified into 6 categories (cut, fast dissolve, fade in, fade out, dissolve, wipe). Different features are used to train different SVMs targeting at different shot boundaries. Researchers explore which features can most effectively classify the shot boundaries.

Shot Boundary Detection with Deep Learning. Hassanien et al. [9] introduces a simple C3D network that takes a segments of fixed length as input and classify it into 3 categories (cut, gradual, background). This method shows the effectiveness of ConvNet in this task. However, this method deals with gradual transitions of different scales in the same way and cannot locate the accurate' boundaries. Gygli [7] also adopts fully convolutional network. It takes the whole video sequence as input and assigns the positive label to the frames in transitions.

Image Similarity Comparison. Deep learning has been successful on image similarity comparison task. In [28], three architectures are proposed to compute image similarities, siamese net, image concatenation net, pseudo-siamese net. Empirical experiments show the image concatenation network and its variants obtain the best performance. In [22], a ranking model that employs deep learning techniques to learn similarity metric directly from images. We apply the similarity measurement only for the cut transition detection.

Object Detection. State-of-the-art methods for general object detection are mainly based on deep ConvNet to extract rich semantic features from images. Liu et al. [15] introduces single shot detector (SSD) using default boxes to match the feature to ground truth and achieve the speed of 19–46 fps. Our gradual detection model design share the same spirit with SSD.

Action Recognition. Carreira and Zisserman [13] has released the kinetics database for large-scale action classification. I3D [3] shows a good weights initialization is necessary to train the C3D network. Qiu et al. [19] proposes a fast network architecture based a spatial convolution kernel and temporal kernel to explore the temporal information. Action recognition is closely related to our work because we want to use temporal information to distinguish large motions and the gradual transitions.

Action Detection. This task focuses on learning how to detect action instances in untrimmed videos. Recently, many approaches adopt detection by classification' framework. Xu et al. [24] builds faster-RCNN style architecture for fast classifying and locating actions. It first selects potential segments with region proposal network and proposes the ROI 3D pooling layer to extract rich features for further classification. In [14], the single shot detector locates action on feature map extracted from well trained action classification ConvNets. Escorcia et al. [6] proposes to generate a set of proposals based on the RNN network. Zhao et al. [30] models the temporal structure of each action instance via a structured temporal pyramid. Although some of the methods can be applied to gradual

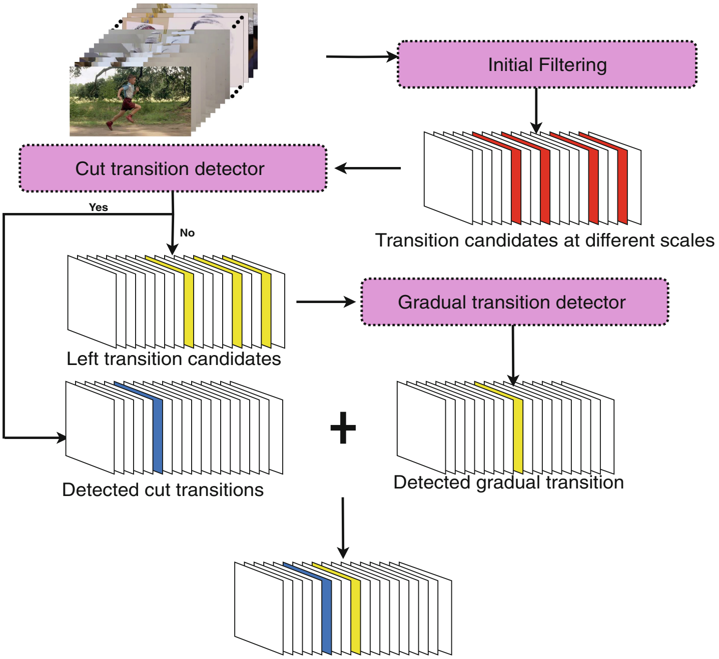

Initial Filtering
Cut transition detector
Transition candidates at different scales
Yes
No
Left transition candidates
Gradual transition detector
+
Detected cut transitions
Detected gradual transition

Fig. 2. An overview of our framework

detection directly, these methods rely on extracting rich spatial-temporal features from a heavy ConvNet body, so these methods are far slower than our proposed methods.

3 Our Approach

In this section, we will introduce our approach in details. The framework of our approach is shown in Fig. 2.

3.1 An Overview

The framework takes a video as input and predicts the locations of transitions. The proposed method, as shown in Fig. 2, is composed of three modules, including initial filtering, cut transition detector and gradual transition detector, implemented with three stages. (1) Adaptive thresholding produces a set of transition candidates. Each candidate comes with a center frame index indicating whether the content in frames has drastic changes. These positions may be transitions or caused by large motion, e.g. camera movement. (2) The candidate transitions are further feed into a strong cut transition detector to filter out false cut transitions. (3) For the remaining center frames which have negative responses to the cut detector, we expand them by x frames on both forward and backward temporal directions to form candidate segments. The gradual transition detector processes all these segments, locating the gradual transitions. The whole framework is designed in a cascade way and the computation of the earlier stage is lighter than the later.

3.2 Initial Filtering

As most of the consecutive video frames are highly similar to each other, a trivial unsupervised algorithm can be applied to reduce the candidate regions for further processing. A fast method, adaptive thresholding, is chosen as the initial filtering step.

Let I_n and I_{n+1} be the potential transition candidates and F_{n-a+1}, F_{n-a+2}, ..., F_{n+a} be a set of features extracted from consecutive video frames in a sliding window of length $2a$ centered at frame n. In practice, we use the feature extracted from SqueezeNet [11] trained on Imagenet [4]. The computation cost in this step is subtle. We calculate the similarity metric of each frame S_i, which is represented as the cosine distance between the current frame feature and its neighboring frame feature. Given the similarity metric of these frames as S_{n-a+1}, S_{n-a+2}, S_{n-a+3}, ..., S_{n+a-1}, the threshold of a window is calculated as

$$T = t + \frac{\sigma}{2\alpha} \sum_{i=n-a+1}^{n+a-1} (1 - S_i) \tag{1}$$

The hyper-parameter σ is the dynamic threshold ratio and t is the static threshold. In practice, we set σ to 0.05 and t to 0.5. The frame is selected as a candidate center if $1 - S_n$ is larger than T. Lengths of gradual transitions vary greatly. In order not to miss any gradual transition, we down-sample frames with multiple temporal scales. At scale ω, we sample one video frame every ω frames and do the above thresholding operations on these down-sampled frames. Finally, results of different scales are merged together. If two candidates on different scales are too close, i.e., within a distance of 5 frames. The candidate with a lower scale will be kept. In practice, we use scales of 1, 2, 4, 8, 16, and 32.

3.3 Cut Model

Some image pairs are semantically similar even when they are cut transitions, i.e. images containing the same object but the backgrounds are different. Therefore, a stronger cut transition detector is needed to filter out these negative cut candidates from the candidates selected by adaptive thresholding. Zagoruyko and Komodakis [28] show CNN can learn the similarity function directly from image pairs. We design a ConvNet to determine whether a image pair is a cut transition or not. In this paper, we compare four models, including siamese, image concatenation, feature concatenation and C3D ConvNet. In contrast to deepSBD, where the position of the cut transition is unknown in one segment, adaptive thresholding can find the cut transition position accurately since it selects the pair of adjacent frames with the largest dissimilarity as the center, facilitating the learning task for our cut detector.

Siamese. A siamese neural network consists of twin networks that accept distinct images and output their features. The parameters are shared between the twin networks and each network computes the same function. An energy loss function is added to the top for optimization. In our problem, we choose contrastive loss as the top energy function. The siamese net outputs a similarity score. At inference, we select the score above some threshold.

Feature Concatenation. This network can be seen as a variant of siamese network. More specifically, it has the structure of the siamese net described above, computing the feature using the same network architecture and weights. The loss energy function is not applied directly to the features. Instead, we concatenate features from both images and add cross entropy loss function to the top.

Image Concatenation. We simply consider the two patches of an RGB image pairs as a 6-channel image and feed it to a generic network. This network provides greater flexibility compared to the above models as it starts by processing the two patches jointly. It is fast to train and infer. Further more, it allows to concatenate multiple images as a input. We find the performance is much improved when using more images.

C3D ConvNet. Hassanien et al. [9] shows the C3D ConvNet is capable of classifying cut transitions. Therefore, we also test this structure for comparison. However, the C3D ConvNet is more complex than 2D ConvNet, which requires much computation resources.

3.4 Gradual Model

Inspired by region proposal network [20] and single shot detector [15], we propose a single shot boundary network, a novel network to locate gradual transitions in a continuous video stream. The network, illustrated in Fig. 3, consists of 2 components, a shared C3D ConvNet feature extractor and subnets for classification and localization.

Feature Hierarchies. Innovated by deepSBD, the C3D ConvNet shows impressive performance in this task. Therefore, we use a C3D ConvNet to extract rich temporal feature hierarchies from a given input video buffer. The input to our model is a sequence of RGB video frames with a dimension of $3 \times L \times H \times W$ and we use ResNet-18 proposed in [8] as the backbone network. We modify all the temporal strides to 1 in ResNet-18 so that the length of the final feature map is also L. The number of frames L can be arbitrary and is only limited by memory.

Subnets for Classification and Location. Since the lengths of gradual transitions are various, we use the same notion default boxes introduced in [15]. In our task, we call it default segments. Default segments are predefined multi-scale

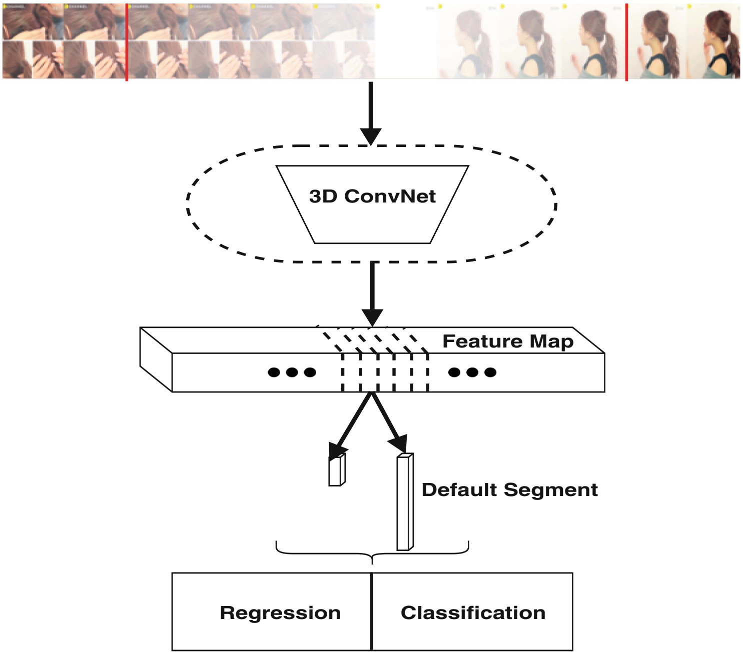


Fig. 3. An overview of gradual detector

windows centered at a location. we put one default segment every $l \times (1 - a)$ frames where l is the length of the default segment and a is the positive IOU threshold. Therefore, each ground truth whose length is between l/a and $l \times a$ can be matched to a default segment. The total number of default segments is $L/(l \times (1 - a))$. The default segments serve as reference segments for ground truth matching. To get features for predicting gradual transitions, we first apply a spatial global average pooling layer to reduce the spatial dimension to 1×1. At each location which has k default segments, we apply a $2k \times 3 \times 1 \times 1$ filter A for binary classification, and a $2k \times 3 \times 1 \times 1$ filter B for location refinement. For both A and B, 3 is the size of the temporal convolution kernel. For A, 2 corresponds to binary classification of a gradual transition or not. For B, 2 corresponds to two relative offsets of $\{\delta c_i, \delta l_i\}$ to the center location and the length of each default segment respectively, where the ground truth of $\{\delta c_i, \delta l_i\}$ is defined as

$$\delta c_i = (c - c_i)/l_i \tag{2}$$

$$\delta l_i = log(l/l_i) \tag{3}$$

The mark c_i and l_i are the center location and the length of default segments while c and l is the ground truth position and length.

Optimization Strategy. In training, positive/negative labels are assigned to default segments. Following the same protocol in object detection, positive labels are assigned if default segments are overlapped with some ground truth if intersection of union $IOU > a$ and negative labels are assigned for default segments if $IOU < b$. Segments with IOU between a and b are ignored during training. In practice, we set a to 0.5 and b to 0.1, which achieves the best performance. As the length of the gradual transitions in our training data ranges in 3 to 40, we use 2 default segments of length 6 and 20 to cover all true transitions. Similar to single shot detector, we implement hard negative example mining and dynamically balance the positive and negative examples with a ratio of 1:1 during training. To utilize the GPU efficiently, we fixed the length of each segment, consisting of L consecutive frames, i.e., L is 64 in our experiment.

We train the network by optimizing the classification and the regression losses jointly with a fixed learning rate of 0.001 for 5 epochs. We adopt softmax loss for classification and smooth L1 loss for regression. The loss function is given in (4). The hyper-parameter λ is set to 1 in practice. Y_i^1 is the predicted score and T_i^1 is the assigned label. $Y_i^2 = \{\delta c_i, \delta l_i\}$ is the predicted relative offset to the default segments and T_i^2 is the target location. The loss function is the same as [15], which is

$$Loss = \frac{1}{N_{cls}} \sum_i L_{cls}(Y_i^1, T_i^1) + \lambda \frac{1}{N_{loc}} \sum_i L_{loc}(Y_i^2, T_i^2) \tag{4}$$

Inference. At inference, the framework processes input videos of varying lengths. However, in order not to exceed the limit of memory, a video will be divided into segments of length T_{seg} with a overlap of $\frac{1}{2}T_{seg}$ such that transitions won't be missed due to the division. After predicting one video, we apply non maximum suppression (NMS) to all the predictions. If two predicted gradual transitions are overlapped, we remove the one with lower classification score.

4 ClipShots

Current datasets, i.e. TRECVID and RAI, are not sufficient for training deep neural network due to a limited size. In addition, previous work utilized different training sets when evaluating their supervised methods on TRECVID and RAI. Therefore, a benchmark is made for comparing different methods fairly. ClipShots is the first large-scale dataset for shot boundary detection collected from Youtube and Weibo covering more than 20 categories, including sports, TV shows, animals, etc. In contrast to TRECVID2007 and RAI, which only consist of documentaries or talk shows where the frames are relatively static, we construct a database containing 4039 short videos from Youtube and Weibo. Many short videos are home-made, with more challenges, e.g. hand-held vibrations and large occlusion. The training set consists of 3539 videos, 122760 cut transitions, and 35698 gradual transitions while the evaluation set consists of 500 videos, 5876 cut transitions, and 2422 gradual transitions. The types of these videos are various, including movie spotlights, competition highlights, family videos recorded by mobile phones etc. Each video has a length of 1–20 min. The gradual transitions in our database include dissolve, fade in fade out, and sliding in sliding out. In order to annotate such a large dataset, we design an annotation tool allowing annotators to watch multiple frames on a single page and select the begin frame and the end frame of transitions. More details are given in the appendix.

5 Experiments

5.1 Databases and Evaluation Metrics

Training and Evaluation Set. The proposed framework is trained and tested on ClipShots. In order to illustrate the effectiveness of our approach and ClipShots, we also evaluated them on two public databases (TRECVID2007, RAI).

Evaluation Metrics. For all 3 databases, we use the standard TRECVID evaluation metrics: one-to-one match if the predicted boundary has at least 1 frame overlapped with the ground truth. For our testing set, we add an additional criterion using IOU to measure the localization performance. We assess performance quantitatively using precision (P), recall (R) and F-score (F).

5.2 Experiments Configuration

We adopt adaptive thresholding to find candidate segments and adjust the parameters to make sure it achieves nearly 100% recall for both cut and gradual transitions. For cut detector, 122760 positive examples and 224312 negative examples are used for training. For gradual detector, the training set contains 35698 ground truths. The potential segments filtered by adaptive thresholding are divided into subsegments of fixed length 64, with overlapped length of 32 between 2 consecutive segments. We choose ResNet-18 3D- ConvNet as the backbone, setting all the strides in the temporal dimension to 1 so that the temporal length of the output feature is identical with the input length. The weights of 3D ResNet-18 are initialized with model pretrained on kinetics database, as the inflated 3D-Conv [3]. For both cut and gradual model, the positive examples and negative examples are highly unbalanced so the positive and negative samples are dynamically balanced with ratio 1:1 in each mini-batch.

5.3 Experiments on ClipShots

Cut Detector Comparison. In this section, we choose four potential models introduced in Sect. 3.3 and test their performance. We use ResNet-50 as the backbone for all models and a fixed learning rate of 0.0001, We train each model for 5 epochs from scratch. For C3D, we adopt the same configuration as deepSBD. For image concatenation model, we evaluated it with different number of images. We expand x frames to the forward and backward in the temporal direction. As Table 1 shows, the image concatenation model obtains best performance among these four models when using 4 or more frames. Siamese net performs worse than image concatenation (2 frames) and C3D network. Given the fact that siamese

Table 1. Comparison of cut models. Image concat(6 frames) obtains the best performance.

Model	P	R	F
Image concat(2 frames)	0.771	0.793	0.782
Image concat(4 frames)	0.775	0.862	0.816
Image concat(6 frames)	0.776	0.934	0.848
Feature concat(2 frames)	0.231	0.574	0.329
Siamese(2 frames)	0.638	0.852	0.729
C3D(16 frames)	0.760	0.910	0.831

net cannot explore information on multiple frames and its computation cost is much larger than image concatenation, this architecture is not adopted in our framework. C3D network (16 frames) is a little better than image concatenation (2 frames), but much worse than image concatenation (4 frames or 6 frames). Feature concatenation is not a working architecture, but we still list it here. For image concatenation, we also study the relationship between the number of input images and performance. More input frames can improve performance. The model gains improvements when increasing the frame number from 2 to 6 and saturates around 6. Therefore, we use an input of 6 frames in our method considering both performance and the processing speed.

Table 2. All methods under a unified viewpoint. Different cut models and gradual models are compared.

Method	Initial filtering	Cut	Gradual
(1)	No	DeepSBD	DeepSBD
(2)	Yes	DeepSBD	DeepSBD
(3)	Yes	Image concat(6 frames)	DeepSBD
(4)	Yes	Image concat(6 frames)	SSBD

Ablation Study. We conduct ablation study with different options. The detailed setting is shown in Table 2. The difference is mainly at cut models, gradual models, and whether initial filtering is used. We also implement deepSBD but the post processing technology introduced in [9] is abandoned for a fair comparison. We adopt 3D ResNet-18 as the backbone for both deepSBD and our single shot boundary detector (Table 3).

Method (1). The model classifies segments directly into 3 categories (cut, gradual, and background).

Method (2). Compared to method (1), initial filtering is utilized to find candidate segments for deepSBD. As is shown, the performance of gradual transition is

Table 3. Performance of different methods. Our method (4) obtains the best performance in both cut transition detection and gradual transitions detection.

Methods	Cut			Gradual		
	P	R	F	P	R	F
(1)	0.765	0.910	0.831	0.770	0.622	0.688
(2)	0.757	0.902	0.823	0.699	0.810	0.750
(3)	0.776	0.934	0.848	0.711	0.830	0.766
(4)	0.776	0.934	0.848	0.840	0.904	0.870

higher than the original deepSBD. It is implied that the initial filtering can also improve performance of deepSBD.

Method (3). For gradual transitions, the deepSBD model only classifies the segments into 2 categories (gradual transition and background) so cut transitions are treated as negative samples. For cut detector, we use image concatenation model. The results show the single shot boundary detector is better than deepSBD by a large margin.

Method (4). The results reveals that our single shot boundary detector is far better than deepSBD. We attribute the performance gain to the following reasons: (1) The receptive field of our model is much bigger than deepSBD, hence the detector can exploit more temporal information. (2) Our default segment design is effective for dealing with gradual transitions of multi scales.

Benchmark in ClipShots. We implement [9] and evaluate them in ClipShots. Table 4 summaries performance of different methods. DeepSBD with 3D ResNet-18 is significantly better than the original network (3D Alexnet alike).

Table 4. Benchmark in ClipShots

Methods	Cut			Gradual		
	P	R	F	P	R	F
deepSBD (Original)	0.731	0.921	0.815	0.837	0.386	0.528
deepSBD (ResNet-18)	0.765	0.910	0.831	0.770	0.622	0.688
DSM (Ours)	0.776	0.934	0.848	0.840	0.904	0.870

Speed Comparison. In this section, we compare the speeds of different models as shown in Table 5. The code is implemented using PyTorch and tested with one TITAN XP GPU. Our method is nearly 2 times faster than the original deepSBD on account of adaptive thresholding based initial filter (Table 6).

Gradual Model Localization Performance. An accurate localization of gradual transitions is important in many video recognition task. Therefore, we

Table 5. Comparison of speed

Model	Speed (FPS)
deepSBD	382
Adaptive thresholding+deepSBD	680
Ours	700

Table 6. Localization performance. We calculate the F1-score at different IOU threshold.

IOU	P	R	F
0.1	0.813	0.865	0.839
0.5	0.726	0.772	0.748
0.75	0.599	0.637	0.618

also evaluate performance of the gradual transition localization using the proposed framework. F1 scores are measured at different IOU level $(0.1, 0.5, 0.75)$. A predicted gradual transition is considered as correct only if its $IOU > a$, otherwise it's considered wrong. When the IOU is 0.75, we can still obtain a F1 score of 0.618, indicating the proposed gradual detector is able to accurately locate gradual transitions.

Table 7. Trecvid07 top performers.

Methods	Cut			Gradual		
	P	R	F1	P	R	F1
ATT [16]	0.996	0.979	0.972	0.802	0.709	0.753
THU11 [26]	0.982	0.968	0.975	0.733	0.718	0.725
Marburg [18]	0.942	0.945	0.944	0.595	0.766	0.670
NHT [12]	0.975	0.816	0.945	0.768	0.578	0.66
Priya [5]	0.972	0.976	0.974	0.869	0.719	0.78
DeepSBD [9]	0.978	0.968	0.973	0.826	0.731	0.776
Ours	0.971	0.988	0.980	0.813	0.806	0.810
Ours (correct label)	0.981	0.997	0.989	0.838	0.845	0.841

5.4 Experiments on TRECVID07

TRECVID07 contains a total of 17 videos, including 2236 cut transitions and 225 gradual transitions. They are all color and black/white documentaries. The videos include cases such as global illumination variation, smoke, fire, and fast non-rigid motion. We take the ground truth from TRECVID07 SBD task. In addition, the experimental results of the proposed method over this database are compared to the top performers of TRECVID07 SBD task. We find some of the ground truths are wrong, so we correct these labels. Evaluation results using original labels and corrected labels are both reported. The cut and gradual models are trained with the same training setting described in Sect. 5.2.

In Table 7, we present a comparative evaluation of the shot boundary detection performance with existing state-of-the-art approaches in terms of F1-score and report the results using both the original ground truth and the corrected ground truth. We evaluate cut transitions and gradual transitions separately. Cut transitions are the most part of all transitions in a video so it plays a dominate role in the overall performance. For cut transitions, we improve the-state-of-art by 0.6%, which is a huge improvement considering there is no much space for improvement. In fact, the errors concentrate in black/white videos due to the lack of similar ones in the training set. Further improvement can be achieved through adding more black/white videos into the training set. For gradual transitions, we achieve 2.9% improvement comparing to the state-of-the-art when using the original ground truth and 6.4% improvement when using the corrected ground-truth (Table 8).

Table 8. RAI comparison

	F1 score
Apostolidis et al. [1]	0.84
Baraldi et al. [2]	0.84
Song et al. [21]	0.68
Michael et al. [7]	0.88
Hassanien et al. [9]	0.934
Ours	0.935

5.5 Experiments on RAI

RAI database is a collection of ten randomly selected broadcasting videos from the Rai Scuola video archive 1, which are mainly documentaries and talk shows. This database includes 722 cut transitions and 263 gradual transitions. Shots have been manually annotated by a set of human experts. The proposed method achieves a competitive results compared to deepSBD. It is noted that DeepSBD adopts posting-processing technology, i.e. filtering the segments whose HSV similarity under a threshold, which is not used in our methods. We perform evaluations on TRECVID and RAI using the same models, weights, and hyperparameters, which indicates the proposed framework are robust on different databases.

6 Conclusion

We propose a cascade shot transition detection framework and annotate the first large-scale shot boundary database. Adaptive thresholding is adopted to find candidate regions for acceleration. The cut and gradual transition detector are designed separately. The cut transition detector is for measuring similarity while the gradual transition detector is for capturing temporal patterns. Especially, the gradual detector is able to locate gradual transitions of multi-scales. We outperform state-of-the-art methods on both TRECVID and RAI databases. In addition, our framework is very fast, achieving a 30× real-time speed.

References

1. Apostolidis, E., Mezaris, V.: Fast shot segmentation combining global and local visual descriptors. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6583–6587. IEEE (2014)
2. Baraldi, L., Grana, C., Cucchiara, R.: Shot and scene detection via hierarchical clustering for re-using broadcast video. In: Azzopardi, G., Petkov, N. (eds.) CAIP 2015, Part I. LNCS, vol. 9256, pp. 801–811. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23192-1_67
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. IEEE (2017)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 248–255. IEEE (2009)
5. Domnic, S.: Walsh-Hadamard transform kernel-based feature vector for shot boundary detection. IEEE Trans. Image Process. **23**(12), 5187–5197 (2014)
6. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: DAPs: deep action proposals for action understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part III. LNCS, vol. 9907, pp. 768–784. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_47
7. Gygli, M.: Ridiculously fast shot boundary detection with fully convolutional neural networks (2017). arXiv preprint: arXiv:1705.08214
8. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? (2017). arXiv preprint: arXiv:1711.09577
9. Hassanien, A., Elgharib, M., Selim, A., Hefeeda, M., Matusik, W.: Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks (2017). arXiv preprint: arXiv:1705.03281
10. Huang, Q., Xiong, Y., Xiong, Y., Zhang, Y., Lin, D.: From trailers to storylines: an efficient way to learn from movies (2018). arXiv preprint: arXiv:1806.05341
11. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5 MB model size (2016). arXiv preprint: arXiv:1602.07360
12. Kawai, Y., Sumiyoshi, H., Yagi, N.: Shot boundary detection at TRECVID 2007. In: TRECVID (2007)
13. Kay, W., et al.: The kinetics human action video dataset (2017). arXiv preprint: arXiv:1705.06950
14. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 2017 ACM on Multimedia Conference, pp. 988–996. ACM (2017)
15. Liu, W., et al.: SSD: single shot MultiBox detector. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part I. LNCS, vol. 9905, pp. 21–37. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_2

16. Liu, Z., Gibbon, D., Zavesky, E., Shahraray, B., Haffner, P.: At&t research at TRECVID 2007. In: Proceedings of TRECVID Workshop, pp. 19–26 (2007)
17. Lu, Z.M., Shi, Y.: Fast video shot boundary detection based on svd and pattern matching. IEEE Trans. Image Process. **22**(12), 5136–5145 (2013)
18. Mühling, M., Ewerth, R., Stadelmann, T., Zöfel, C., Shi, B., Freisleben, B.: University of Marburg at TRECVID 2007: shot boundary detection and high level feature extraction. In: TRECVID (2007)
19. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 5534–5542. IEEE (2017)
20. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
21. Song, Y., Redi, M., Vallmitjana, J., Jaimes, A.: 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, pp. 659–668. ACM (2016)
22. Wang, J., et al.: Learning fine-grained image similarity with deep ranking (2014). arXiv preprint: arXiv:1404.4661
23. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: UntrimmedNets for weakly supervised action recognition and detection. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2 (2017)
24. Xu, H., Das, A., Saenko, K.: R-C3D: region convolutional 3D network for temporal activity detection. In: The IEEE International Conference on Computer Vision (ICCV), vol. 6, p. 8 (2017)
25. Yuan, J., Li, J., Lin, F., Zhang, B.: A unified shot boundary detection framework based on graph partition model. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 539–542. ACM (2005)
26. Yuan, J., et al.: A formal study of shot boundary detection. IEEE Trans. Circ. Syst. Video Technol. **17**(2), 168–186 (2007)
27. Yusoff, Y., Christmas, W.J., Kittler, J.: Video shot cut detection using adaptive thresholding. In: BMVC, pp. 1–10 (2000)
28. Zagoruyko, S., Komodakis, N.: Learning to compare image patches via convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4353–4361. IEEE (2015)
29. Zhang, K., Chao, W.-L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016, Part VII. LNCS, vol. 9911, pp. 766–782. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46478-7_47
30. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: The IEEE International Conference on Computer Vision (ICCV), vol. 8 (2017)