Pentuplet Loss for Simultaneous Shots and Critical Points Detection in a Video

Nitin Gupta, Abhinav Jain, Prerna Agarwal, Shashank Mujumdar, Sameep Mehta
IBM Research, India
{ngupta47, abhinavj, prernaagarwal, shamujum, sameepmehta}@in.ibm.com

***Abstract*—Critical events in videos amount to the set of frames where the user attention is heightened. Such events are usually fine-grained activities and do not necessarily have defined shot boundaries. Traditional approaches to the task of Shot Boundary Detection (SBD) in videos perform frame-level classification to obtain shot boundaries and fail to identify the critical shots in the video. We model the problem of identifying critical frames and shot boundaries in a video as learning an image frame similarity metric where the distance relationships between different types of video frames are modeled. We propose a novel *pentuplet loss* to learn the frame image similarity metric through a pentuplet-based deep learning framework. We showcase the results of our proposed framework on soccer highlight videos against state-of-the-art baselines and significantly outperform them for the task of shot boundary detection. The proposed framework shows promising results for the task of critical frame detection against human annotations on soccer highlight videos.**

I. Introduction

In recent years, due to the increase in consumption of digital media, particularly in video data, automatic and intelligent processing of video data has become necessary for gathering insights into its temporal and spatial events. A significant subset of the digital media is consumed in the form of highlight videos. Highlights are short (typically 2-4 minutes long) duration recaps of interesting events happened during an event that are quick to render. These events are shown in the form of shots/clips which are series of interrelated consecutive pictures taken contiguously representing a continuous action in time and space, with variability in angle and location of the film. These videos are used as in marketing the actual event and often monetized with sponsor content. In such a scenario, it becomes increasingly important to identify the critical events in the video that engage the viewer. In this paper, we focus on the task of identifying critical events from a highlight video and simultaneously propose a framework for accurate shot boundary detection (SBD).

There has been a strong focus on the task of SBD in the literature which is an important pre-processing step to enable video analysis. Traditionally, methods have relied on hand-crafted features [1]–[4], spatial information and intensity histograms [3]–[5] to detect shot boundaries. In [3] a fast algorithm to detect cut transition (abrupt changes) and gradual transitions is developed. In [2] a method is proposed that uses wavelet-based feature vector which measures color, edge and motion strength. The feature vector is extracted for each frame and frame difference is used to identify temporal changes. Although these methods are fast, they compromise on the detection accuracy. Deep learning based solutions have been proposed recently, [6], [7] to improve the detection accuracy and learn to detect sharp as well as gradual transitions within videos. Open source libraries have also been developed that detect shots and scenes from a video [8], [9]. However, a glaring drawback of these approaches is their inability to capture fine-grained critical events within the video. Critical events can be defined as fine-grained activities that grab the user attention that tend to increase user anxiety and anticipation. These activities do not necessarily have defined shot boundaries and could be part of a whole video event. Clip 2 shown in Fig. 2 illustrates an example of a critical event in a soccer highlight video. Other examples of critical events seen generally in videos are - (i) reality show result about to be announced, (ii) a doctor about to break outcome of surgery, (iii) a criminal about to commit a murder etc. Visually, the frames that depict the critical event are similar to those of the overall event they belong to, which results in poor classification accuracy of the conventional SBD methods. Thus, instead of learning a classifier to accurately label the frames in the video, we introduce the *pentuplet loss* that is able to learn a frame similarity metric by modeling the distance relationships between different types of frames in a video.

Formally, the pentuplet-loss function considers a set of four image frames: positive (p), intermediate-1 (i^1), intermediate-2 (i^2) and negative (n), for a given query image (q), and learns the similarity metric between these images. If $dist(a,b)$ denotes the distance between two images (a,b), then the pentuplet-loss models the relation: $dist(q,p) < dist(q,n)$ and $dist(p,i^1) < dist(p,i^2) < dist(p,n)$. Here, p is sampled from the same clip as q and is not a critical frame while i^1 and i^2 are sampled from the set of critical frames within the same clip as q. Specifically, i^1 is sampled from initial frames of critical clip and i^2 from final frames. Such fine grained distinction breaks down a critical event into sub-events which have traits that separate them from each other and from rest of the clip and traits shared amongst them. Our dataset have annotations as shown in Fig.1 for which details are provided in Sec. IV-C2. Lastly, n is sampled from the next consecutive clip in the video. The implicit relationships between the images are controlled by four separate margins in the pentuplet loss. Using the learned similarity metric, we empirically determine the threshold on the L2 norms of the different image frame pairs to identify critical frames and shot boundaries.

We showcase the results of our proposed framework on soccer highlight videos. They usually contain goals scored and missed, fouls, a celebration by players, crowd's visual response etc. The shots are juxtaposed using abrupt transitions made by hard cuts where one frame belongs to the first clip and the next frame belongs to the second shot. We observe the following types in the soccer highlight videos: (i) 'close up shots' where the subject fills the field of vision on the camera for example in celebration clips, (ii) 'long shots' where objects are far removed from the camera providing a complete picture of the event happening, (iii) 'medium shots' that lie between the extremes of long and close up shots and (iv) 'transitional shots' where the clip covers an event with a gradual smooth transition between primary shot types. These different types of shots contain a variety of critical scenes and serve as a good use case for both the task of shot boundary and critical frame detection. Although, we obtain annotations and present results on a dataset of soccer highlight videos, the proposed framework is generic and can easily be adapted to any other type of videos.

In summary, the main contributions of this work are as follows:

1) We introduce the novel pentuplet-loss that learns an image similarity metric to simultaneously detect the critical frames and shot boundaries in a given video.
2) We create a dataset of soccer highlight videos with human annotations on 700 clips with 11,470 frames.
3) For the novel task of critical frame identification, we report the accuracy of our framework using human annotated labels.
4) We introduce a generic supervised deep learning based framework to simultaneously detect critical frames and shot boundaries from videos.

II. Proposed Method

A pentuplet-based network is proposed to detect the shot boundaries (clips), as well as a critical portion of detected shots, if exists. This network takes a pentuplet as input. A pentuplet contains a query image (q), positive image (p), intermediate image-1 (i^1), intermediate image-2 (i^2), and negative image (n). The intermediate image-1 and 2 refer to samples, that does not qualify as positive samples but is nearer to both the query and positive sample in terms of relevance. These pentuplets are selected from training samples and fed into five identical deep neural networks (Alexnet [10]), with shared architecture and parameters. The output of these networks is fed into L2-normalization layer, which helps to normalize the features to have unit L2 norm.

A pentuplet loss ranking layer, over the top of fully connected layer, evaluates the proposed pentuplet loss over the set of training pentuplets. During learning, it tries to preserve the relative order across the samples in selected pentuplet (query (q), positive (p), intermediate-1 (i^1), intermediate-2 (i^2), and negative (n)). The main goal is to find the feature embedding such that: (a) positive and negative samples lies in two different embedding space, which helps to detect the shots boundaries in a video; and (b) intermediate samples lies near to positive samples with some margin which is smaller to the margin between positive and negative samples, which helps to detect critical scene within a detected shots (if exist). Whenever there is a violation in the ordering of pentuplets, the loss is calculated and the gradients are computed over the loss and backpropagated to the lower layers. The lower layers adjust their parameters to minimize the proposed pentuplet loss to learn feature embedding space, which helps to simultaneously detect shots as well as critical points within a detected shots.

A. Pentuplet loss function and its optimization

The pentuplet loss considers four samples from the database: Positive ($p_k \in P$), Intermediate-1 ($i_k^1 \in I_1$), Intermediate-2 ($i_k^2 \in I_2$) and Negative ($n_k \in N$) for each of the query sample $q_k \in Q$. Here P, I_1, I_2, N denote the sets containing positive, intermediate-1, intermediate-2, negative samples respectively for the queries belonging to the set Q. Our goal is to learn an embedding function $f(.)$ such that following relations (see Eq. 1 to 3) exists across samples, which helps to simultaneously detect shots as well as critical points within detected shots:

$$dist(f(q_k), f(p_k)) < dist(f(q_k), f(n_k)) \tag{1}$$

$$\text{and } dist(f(p_k), f(i_k^1)) < dist(f(p_k), f(i_k^2)) \tag{2}$$

$$< dist(f(p_k), f(n_k)) \tag{3}$$

$$\text{where, } dist(f(a), f(b)) = ||f(a) - f(b)||_2^2$$

$$\text{and (a, b)} \in \{q_k, p_k, i_k^1, i_k^2, n_k\}$$

$\{q_k, p_k, i_k^1, i_k^2, n_k\}$ is the set representing the k^{th} pentuplet used during the training phase to estimate the optimal embedded space. Pentuplets are selected using the hard mining technique proposed by [11]. Based on the properties of pentuplet loss as given in Eq. 1 to 3, the following equations can be written:

$$||f(q_k) - f(p_k)||_2^2 - ||f(q_k) - f(n_k)||_2^2 \quad < \quad 0 \tag{4}$$

$$||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k - f(n_k)||_2^2 \quad < \quad 0 \tag{5}$$

$$||f(p_k) - f(i_k^2)||_2^2 - ||f(p_k) - f(n_k)||_2^2 \quad < \quad 0 \tag{6}$$

$$||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(i_k^2)||_2^2 \quad < \quad 0 \tag{7}$$

We have also considered hinge function to model the losses as discussed in [12] given as.

$$\mathcal{L}_1(q_k, p_k, n_k) = max\{0, g_1 + ||f(q_k) - f(p_k)||_2^2 - ||f(q_k) - f(n_k)||_2^2\} \tag{8}$$

$$\mathcal{L}_2(p_k, i_k^1, n_k) = max\{0, g_2 + ||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} \tag{9}$$

$$\mathcal{L}_3(p_k, i_k^2, n_k) = max\{0, g_3 + ||f(p_k) - f(i_k^2)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} \tag{10}$$

$$\mathcal{L}_4(p_k, i_k^1, i_k^2) = max\{0, g_4 + ||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(i_k^2)||_2^2\} \tag{11}$$

Here, g_1, g_2, g_3 and g_4 denote the gap parameters that regularizes the gaps between the image pairs $\{(q_k, p_k)$ and $(q_k, n_k)\}$, $\{(p_k, i_k^1)$ and $(p_k, n_k)\}$, $\{(p_k, i_k^2)$ and $(p_k, n_k)\}$,

$\{(p_k, i_k^1)$ and $(p_k, i_k^2)\}$ respectively. Finally, by using Eq. 8, 9, 10 and 11, we can formulate our constrained optimization problem as:

$$\text{minimize} \sum_k \xi_k + \sum_k \zeta_k + \sum_k \varrho_k + \sum_k \Delta_k + \lambda ||W||_2^2 \tag{12}$$

subjected to:

$$max\{0, g_1 + ||f(q_k) - f(p_k)||_2^2 - ||f(q_k) - f(n_k)||_2^2\} \leq \xi_k$$
$$max\{0, g_2 + ||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} \leq \zeta_k$$
$$max\{0, g_3 + ||f(p_k) - f(i_k^2)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} \leq \varrho_k$$
$$max\{0, g_4 + ||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(i_k^2)||_2^2\} \leq \Delta_k$$

where, W is the embedding learned by the CNN, and $f(.)$ denotes the function which projects a sample to the embedded space. λ is the regularization parameter that improves the generalization capability of the ranking algorithm. We use $\lambda = .001$ in this paper. The above-constrained optimization problem can be converted into the following unconstrained optimization problem as:

$$\text{minimize} \;\; \lambda||W||_2^2 + \sum_k \mathcal{L}_1(q_k, p_k, n_k) + \sum_k \mathcal{L}_2(p_k, i_k^1, n_k) + \sum_k \mathcal{L}_3(p_k, i_k^2, n_k) + \sum_k \mathcal{L}_4(p_k, i_k^1, i_k^2) \tag{13}$$

$$\text{minimize} \;\; \lambda||W||_2^2 + \sum_k max\{0, g_1 + ||f(q_k) - f(p_k)||_2^2 - ||f(q_k) - f(n_k)||_2^2\} + \sum_k max\{0, g_2 + ||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} + \sum_k max\{0, g_3 + ||f(p_k) - f(i_k^2)||_2^2 - ||f(p_k) - f(n_k)||_2^2\} + \sum_k max\{0, g_4 + ||f(p_k) - f(i_k^1)||_2^2 - ||f(p_k) - f(i_k^2)||_2^2\} \tag{14}$$

When the embedding of the images is normalized to have unit L2 norm, then the proposed optimization function can be converted into a more simpler form as:

$$\text{minimize} \;\; \lambda||W||_2^2 + \sum_k max\{0, g_1 - 2f(q_k)f(p_k) + 2f(q_k)f(n_k)\} + \sum_k max\{0, g_2 - 2f(p_k)f(i_k^1) + 2f(p_k) f(n_k)\} + \sum_k max\{0, g_3 - 2f(p_k)f(i_k^2) + 2f(p_k)f(n_k)\} + \sum_k max\{0, g_4 - 2f(p_k)f(i_k^1) + 2f(p_k)f(i_k^2)\} \tag{15}$$

We use Caffe's inbuilt gradient descent algorithms for solving the unconstrained optimization problem (Eq. 15). As mentioned in [12], the image embedding function $f(.)$ is the composition of the different layers of the deep network model, which is given as: $f(.) = g_n(g_{n-1}(\ldots g_l(.) \ldots))$, where $g_l(.)$ is the forward transfer function of the l^{th} layer. The gradient $\frac{\delta f(.)}{\delta w_l}$ represented as: $\frac{\delta f(.)}{\delta g_l} \times \frac{\delta g_l}{\delta w_l}$, where w_l is the parameter of the l^{th} layer of the deep architecture. Now, $\frac{\delta f(.)}{\delta g_l}$ can be calculated in an iterative fashion as: $\frac{\delta f(.)}{\delta g_{l+1}} \times \frac{\delta g_{l+1}}{\delta g_l}$. Hence, we only need to compute $\frac{\delta g_l}{\delta w_l}$ and $\frac{\delta g_l}{\delta g_{l-1}}$ for the function $g_l(.)$. The ranking layer does not contain any extra parameter and the derivative of the overall loss function $\mathcal{L}$ from Eq. 13 and 15 with respect to the element in a pentuplet set, $f(p_k)$, can be written as:

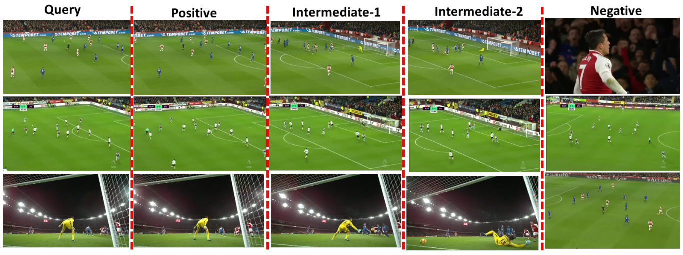

Fig. 1: Pentuplets generated from soccer highlight dataset.

$$\frac{\delta \mathcal{L}}{\delta f(p)} = \begin{cases} 0, & \mathcal{L}_{1-4} = 0 \\ 2(2f(n) - f(q) - f(i^1) - f(i^2)), & \mathcal{L}_{1-3} > 0, \mathcal{L}_4 \leq 0 \\ 2(f(n) - f(q) - 2f(i^1) + f(i^2)), & \\ & \mathcal{L}_{1-2}, \mathcal{L}_4 > 0; \mathcal{L}_3 \leq 0 \\ 2(f(n) - f(i^1) - f(q)), & \mathcal{L}_1, \mathcal{L}_{3-4} > 0; \mathcal{L}_2 \leq 0 \\ 2(2f(n) - 2f(i^1) + f(i^2)), & \mathcal{L}_{2-4} \leq 0; \mathcal{L}_1 > 0 \\ 2(f(n) - f(i^1) - f(q)), & \mathcal{L}_{1-2} > 0; \mathcal{L}_{3-4} \leq 0 \\ 2(f(n) - f(i^2) - f(q)), & \mathcal{L}_1, \mathcal{L}_3 > 0; \mathcal{L}_2, \mathcal{L}_4 \leq 0 \\ 2(f(i^2) - f(i^1) - f(q)), & \mathcal{L}_1, \mathcal{L}_4 > 0; \mathcal{L}_{2-3} \leq 0 \\ 2(2f(n) - f(i^1) - f(i^2)), & \mathcal{L}_{2-3} > 0; \mathcal{L}_1, \mathcal{L}_4 \leq 0 \\ 2(f(n) - 2f(i^1) - f(i^2)), & \mathcal{L}_2, \mathcal{L}_4 > 0; \mathcal{L}_1, \mathcal{L}_3 \leq 0 \\ 2(f(n) - f(i^1)), & \mathcal{L}_{3-4} > 0; \mathcal{L}_{1-2} \leq 0 \\ -2f(q), & \mathcal{L}_1 > 0; \mathcal{L}_{2-4}, \leq 0 \\ 2(f(n) - f(i^1)), & \mathcal{L}_2 > 0; \mathcal{L}_1, \mathcal{L}_{3-4}, \leq 0 \\ 2(f(n) - f(i^2)), & \mathcal{L}_3 > 0; \mathcal{L}_{1-2}, \mathcal{L}_4, \leq 0 \\ 2(f(i^2) - f(i^1)), & \mathcal{L}_4 > 0; \mathcal{L}_{1-3}, \leq 0 \\ 2(2f(n) - f(q) - 2(i^1)), & \mathcal{L}_{1-4} > 0 \end{cases}$$

Similarly, we can calculate derivative of the loss function $\mathcal{L}$ with respect to the other elements in a pentuplet set i.e. $f(q_k)$, $f(i_k^1)$, $f(i_k^2)$ and $f(n_k)$. The back propagation estimates the embedding function $f(.)$ in such a way that the relationships of the pentuplet loss Eq. 1 to 3 are best preserved.

III. Dataset

A. Acquisition Of Dataset

Our Dataset consists of 25 annotated YouTube Soccer Videos, which are highlights of club soccer that are conceptually similar but visually significantly different. Conceptual similarity is dictated by the rules of the game "football" which are universal; on the other hand visual difference is judged by many diverse variables like lighting conditions during the game, professionalism in the game like club or varsity level, camera view point with which video was shot etc. We have collected videos which are professionally played

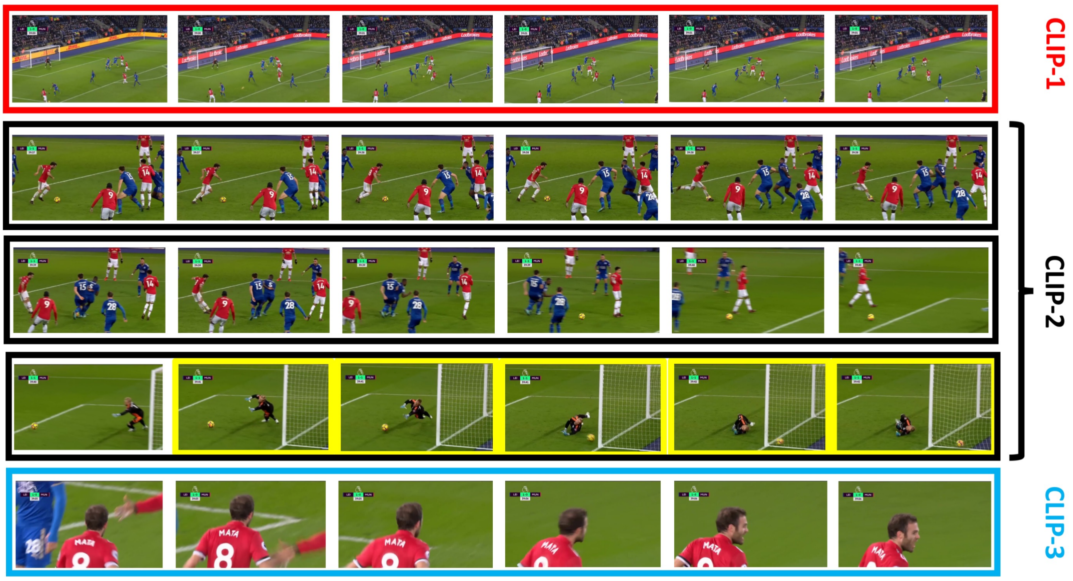

Fig. 2: A sample of temporally annotated continuous clips segmented based on the hard cuts. Clip-1 is a long range shot of ball being passed by one player to some other player. Clip-2 is a medium shot which shows a player in possession of the ball kicking it towards the goal. The highlighted frames of Clip-2 are the critical frames where the ball crosses the goalkeeper and enters the net. Clip-3 is a close up shot which includes all the frames pertaining to celebration of goal.

by trained players while allowing some variability in other factors such as lighting conditions and teams playing to make the learned model more robust and prevent any bias towards the behavior of one of the aforementioned factors. Specifically, the dataset has videos of soccer matches played in following weathers dictating the lighting conditions: daytime, night, bright sunlight and rainy. Also, teams playing the matches belong to different clubs, thus providing variety in color of jerseys and players playing the matches.

To make the task of shot segmentation and simultaneous critical point detection more challenging, we deliberately opted for highlights of soccer videos which have multi-view camera shots of events like goals scored/missed, fouls committed etc.

B. Temporal Segmentation and Dataset Annotation

Highlights are annotated based on the duration of shots/clips where a shot contains all the temporally contiguous frames as shown in Fig.2. A shot boundary is decided based on the presence or absence of a hard cut. In accordance with the visual assessment of shot types, we provide additional annotation with respect to the closeness of the camera shot. For every clip, if within a clip there is a significant difference between consecutive frames based on the closeness of camera shot to the subject, the clip is labeled as "transitional". Otherwise, it is either labeled as "close" if camera shot is a close-up shot or "long" if camera shot is a long-shot.

For critical point detection, based on the definition of criticality we label those frames as "critical" that correspond to the event when a ball is too close to the goal. Due to the inherent subjectivity in the task, it is infeasible to obtain clear-cut ground truth labels for critical frames; thus annotation should be carried out using human judgments. Our approach involves asking 10 users to identify those frames in which the ball gets too close to goalkeeper so much so that it creates high levels of anxiety in them. This often happens when a goal is being scored or closely missed. But amongst users, disagreement in the annotation is observed for the frames which mark the start of a critical clip. Hence, a frame is included in the critical clip only if it receives the same annotation from more than 5 users otherwise it is discarded. Also, all the frames corresponding to the ball passing the goalkeeper and staying inside the net are also labeled as critical. This is done to make sure that critical frames that will act as intermediates to train our network have certain traits that separate them from the rest of the frames in the clip such as closeness of the ball to the goal which our network would learn.

C. Dataset Details

Each highlight video is approximately 2 minutes in length. A total of 700 clips are annotated which have 11,470 frames in total that are extracted at 4 frames per second. 3,500 frames are labeled as critical from 125 responses (5 per video).

IV. Results and Evaluation

A. Evaluation Metrics

We will use the following metrics to measure the quality of clip segmentation:

$$Precision\ (P) = \frac{TP}{TP + FP} \tag{16}$$

$$Recall\ (R) = \frac{TP}{TP + FN} \qquad (17)$$

$$F - Measure\ (F1) = \frac{2 * P * R}{P + R} \qquad (18)$$

where, TP, true positives are correctly marked cuts, FP, false positives are the number of not detected cuts and FN, false negatives are number of falsely detected cuts. Precision is the probability that an assumed cut by the model is actually a cut while Recall is the fraction of existing cuts detected by the model. We will use similar measures to assess the quality of critical frames detection.

B. Baselines

The three baseline methods considered here for comparison are PySceneDetect [8], ShotDetect [9] and state-of-the-art SVD based Shot Boundary Detection [3].

- In *PySceneDetect*, content-aware based detection is used which identifies each location where a fast cut occurs.
- *ShotDetect* is also content based detection method which identifies the shots in a video where transition occurs.
- *SVD-Shot-Detection* first identifies candidate segments to reduce the computational cost of the algorithm. From a set of candidate segments, both cut transitions and gradual transitions are identified. This method use histogram and intensity to calculate frame similarity in HSV color space.

C. Experimental Settings

We have used 250 clips for training and 450 clips for testing. Model is trained for 50,000 iterations with a batch size of 5.

1) Pre-Processing: Laplacian-based approach [13] has been used to remove the blurred frames from both training and testing set. The blurriness of an image can be determined using the concept of variance. A well-focused image is expected to have a high variation (a wide spread of responses, both edge-like and non-edge like). However, very low variance (a tiny spread of responses) indicates very fewer edges in the image, which in-turn implies high blur in an image.

2) Training Configuration Details: For simultaneous learning of shots and critical points within them using proposed pentuplet loss, we have used the AlexNet Model. There is no parameter for the norm layer that we have used on the top of AlexNet. For fully connected layer, we have used: 512 filters, Xavier as weight filler, learning rate multiplier = 10 (as recommended for fine-tuning task) and decay rate multiplier = 1. For pentuplet learning task, we have used the following margin parameters - $g_1 = 0.7$, $g_2 = 0.5$, $g_3 = 0.4$ and $g_4 = 0.1$, determined experimentally. For training the proposed pentuplet network, the pentuplets are generated randomly at each iteration and controlled by the concept of hard negative mining as discussed in [11]. For football dataset, the intermediate-1 samples correspond to the sub-event when ball kicked towards goal is close to it and intermediate-2 samples correspond to the frames where the ball enters the net as shown in Figure 1. Intermediate-1 and 2 samples define the critical points/frames within a detected clip, if they exists. For clips with no critical frames, losses corresponding to intermediate frames are set to 0.

3) Testing: At the time of testing, for the task of clip segmentation, consecutive pair of frames are compared based on the L2 norm of the difference between their corresponding feature vectors; if the L2 norm crosses a threshold value of “2.5”, a hard cut is added between them. Selection for the threshold value is done based on an experimental analysis. Thus, a sequence of frames between two consecutive hard cuts are recognized as one single clip.

In the case of critical point detection, a clip is first detected using above mentioned approach and then it is tested for the presence of critical frames. But here all the frames in a clip are compared with the query frame which is the first frame of the clip. Since, pentuplet network is trained in such a way that intermediates-1 which are also temporally next to positive frames have their feature vectors closer to positive frames' feature vectors in embedding space than intermediates-2. As a result, we identify an increasing or a decreasing trend in L2 norm values which indicates the existence of critical frames in a clip. The trend type depends on the contents of the query frame. Its existence is confirmed by empirical analysis of clips with and without critical frames. However, the frames which are part of the trend are labeled as critical only if the trend holds for at least 8 frames or more which also has been empirically tested.

D. Results and Discussion

Soccer video highlights are more challenging for several reasons: (a) there are rapid movements of camera when the shot is a close up shot or a medium shot; (b) they also possess long shots where there is a great deal of movement from the players, but fairly constrained camera movements; (c) much of background is green (the football field) so that even when shots change, the background is still dominated by a single color. Thus, model learned should be robust to these challenges. Our proposed model overcoming these challenges has the highest precision and recall of all the models (Table I).

Model	Precision(%)	Recall(%)	F1(%)
Proposed Model	**86.09**	**98.02**	**91.67**
PySceneDetect	68.12	74.26	71.09
SVD-SBD	28.65	80.69	42.28
ShotDetect	34.96	81.68	48.96

TABLE I: Performance of models in clip segmentation

Low precision in baselines is due to the excess number of false positives that are detected by them mostly in medium shots due to significant scene changes in a single clip as shown in Fig.2(b). Similarly, higher recall of our proposed model can be explained by cases where baseline models fail to identify the existence of a hard cut between two visually similar frames belonging to two separate temporally contiguous clips. This usually happens with two juxtaposed long shots as shown in Fig.2(a). But, our model segments clips based on the similarity

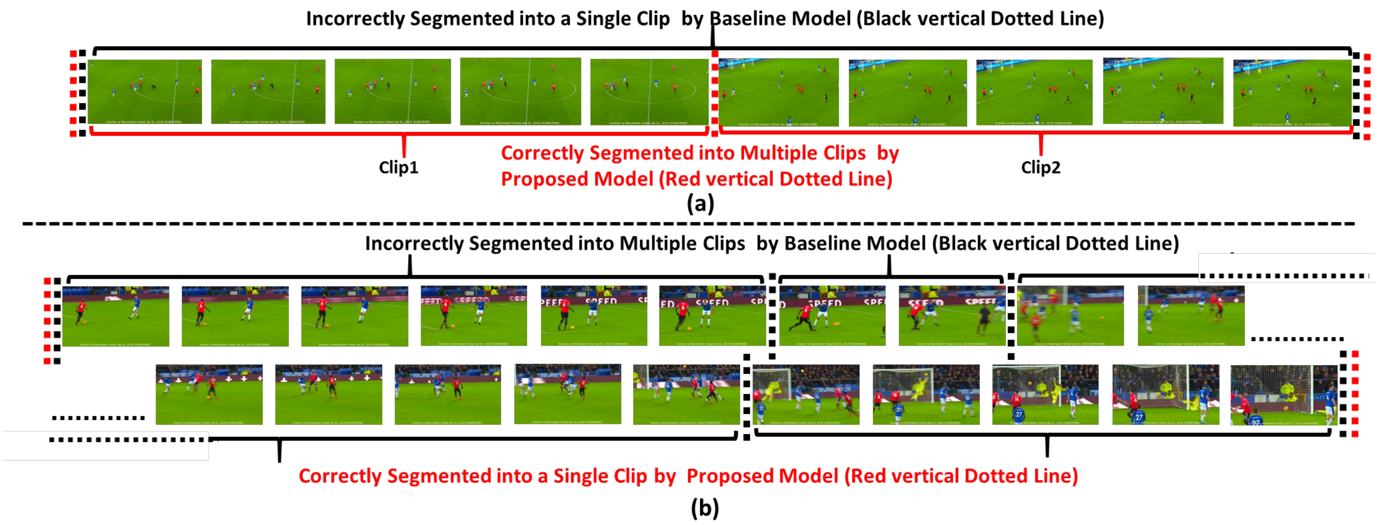
Incorrectly Segmented into a Single Clip by Baseline Model (Black vertical Dotted Line)
Clip1
Correctly Segmented into Multiple Clips by Proposed Model (Red vertical Dotted Line)
Clip2
(a)
Incorrectly Segmented into Multiple Clips by Baseline Model (Black vertical Dotted Line)
Correctly Segmented into a Single Clip by Proposed Model (Red vertical Dotted Line)
(b)

Fig. 3: Two samples from test dataset showcasing the cases where proposed model performs better than the base line models. (a) Proposed model correctly segments the shown frames into two separate clips whereas base line models such as PySceneDetect fails to recognize the presence of hard cut between last frame of clip-1 and first frame of clip-2 clubbing them into a single clip. (b) Example shown here consists of frames that belong to a single clip which are correctly recognized by our proposed model while base line models incorrectly segments them into separate clips marked by black vertical lines.

of the consecutive frames starting from the query frame; thus, with the learned embedding space, it is correctly able to retrieve all the hard cuts which are evident in its higher recall.

Precision(%)	Recall(%)	F1(%)
86.67	78.00	82.11

TABLE II: Performance of our model in critical point detection

For critical frames detection, results are shown in Table II. The analysis shows that false positives correspond to those clips which are close up shots of the ball moving in front of goal but does not correspond to the event where the ball is actually hit towards the goal. So, looking at the frames in isolation, it appears like a goal is being scored; thus, our proposed model confuses them with true positives. Nevertheless, the high F1 value indicates that model is robust in detecting critical frames.

V. Conclusion and Future Work

Clip segmentation is a decade old task but state-of-the-art approaches still struggle when tested on challenging datasets such as "highlights" which we have dealt with in this paper. We proposed a novel framework that can simultaneously do clip segmentation and critical frames detection. The pentuplet network uses a loss function that preserves fine-grained image similarity to learn feature embeddings. The proposed methodology outperforms baseline models in clip segmentation and also gives high performance in the task of critical frame detection which itself is a novel task that has never been targeted before. Critical frames detection might seem a simple binary labeling task, but it's subjectively open to various interpretations. We restricted them by focusing on one domain; thus, introducing a hint of objectivity in the task. Also, the uneven distribution of labeled frames which might make any model biased doesn't affect our model since it is based on similarity learning. In the future, we plan on expanding the domain of critical point detection from soccer videos along with incorporating a temporal aspect of videos in the framework to further enhance the performance.

References

[1] A. F. Smeaton, P. Over, and A. R. Doherty, "Video shot boundary detection: Seven years of trecvid activity," *Computer Vision Image Understanding*, vol. 114, no. 4, 2010.

[2] L. P. G. G. and D. S., "Walsh hadamard transform kernel-based feature vector for shot boundary detection," *IEEE Transactions on Image Processing*, vol. 23, no. 12, 2014.

[3] Z. M. Lu and Y. Shi, "Fast video shot boundary detection based on svd and pattern matching," *IEEE Transactions on Image Processing*, vol. 22, no. 12, 2013.

[4] C. Zhang and W. Wang, "A robust and efficient shot boundary detection approach based on fisher criterion," in *ICM*, 2012.

[5] H. Fang, J. Jiang, and Y. Feng, "A fuzzy logic approach for detection of video shot boundaries," *Pattern Recognition*, vol. 39, no. 11, 2006.

[6] A. Hassanien, M. A. Elgharib, A. Selim, M. Hefeeda, and W. Matusik, "Large-scale, fast and accurate shot boundary detection through spatio-temporal convolutional neural networks," *CoRR*, 2017.

[7] M. Gygli, "Ridiculously fast shot boundary detection with fully convolutional neural networks," *CoRR*, 2017.

[8] "Pyscenedetect," https://pyscenedetect.readthedocs.io/en/latest/.

[9] "Shot detect," https://github.com/johmathe/Shotdetect.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[11] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[12] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, 2014.

[13] J. L. Pech-Pacheco, G. Cristóbal, J. Chamorro-Martinez, and J. Fernández-Valdivia, "Diatom autofocusing in brightfield microscopy: a comparative study," in *ICPR*, 2000.