MUSES

Multi-shot Temporal Event Localization : a Benchmark

Abstract

현재의 temporal event 혹은 action localization에서는 대체로 싱글 카메라의 상황에서 발전해왔다. 하지만 많은 다양한 상황 속에서 이벤트나 행동은 많은 카메라와 다양한 각도에서 포착되어진다. 이 논문에서 우리는 multi-shot temporal event localization 테스크를 제시하며, 이에 따라 MUlti-Shot EventS(MUSES)라는 데이터셋을 제시한다. MUSES는 716시간의 비디오 영상에서 31,477개의 이벤트 객체를 가지고 있다. MUSES의 핵심은 빈번한 짧은 컷인데 이벤트 당 평균 19개, 비디오 당 평균 176개의 샷을 이용하여 큰 변화를 유도한다. 우리는 기존의 Sota model을 이용해 평가해봤을 때 13.1% mAP점수를 기록하였다. 그리고 우리는 18.9% mAP를 기록한 baseline을 제시하낟.

Introduction

수많은 비디오가 생성되고, 공유되고, 소비됨에 따라 최근 들어 video understanding은 많은 관심을 받고 있다. 그 중 하나인 Temporal action localization은 action을 탐지하고 action의 start time과 end time을 예측하는 문제이다. 이러한 task는 보안 감시, 홈 케어, 스포츠 분석 등 다양한 분야에서 매우 중요한 역할을 한다.

이 분야에서 많은 성과가 있었지만, 우리는 이때까지 집중받지 못했던 Tv프로그램과 영화 속 이벤트와 같은 곳에서의 localization을 하고자 한다. 우리는 이것을 multi-shot temporal event localization이라 부르기로 했다.

Motivation

우리는 TV show나 영화에서 생기는 비디오에서 영감을 받았다. 예를 들어, 트레일러는 관객들을 놀려주는 하이라이트를 보여주기 위해 생성되고, 요약본은 관객들이 중요한 서사나 캐릭터에 대한 이해를 돕기 위해 만들어지고, 매시업은 같은 테마의 여러 클립이 혼합되어 제작되어집니다. 이를 위해, 기본적이지만 영상을 편집하기 위한 시간 소모가 많습니다. 그래서 우리는 multi-shot temporal event localization이 이벤트 추출과 비디오 콘텐츠 생성에 많은 도움을 줄 것이라 생각했습니다.

Characteristic

사용자가 만든 비디오나 감시 카메라와 비교해보면 tv 쇼와 영화는 shot cuts이 많습니다. shot의 의미는 한대의 카메라가 중단 없이 촬영한 것입니다. multi-camera와 편집 기술 때문에 action or event는 여러 shot의 모음으로 연결되어집니다. 예를 들어, cutting, dissolve, cut-in, cross-cut과 같은. 다른 말로 shot의 종료는 action의 종료가 아님을 의미합니다.

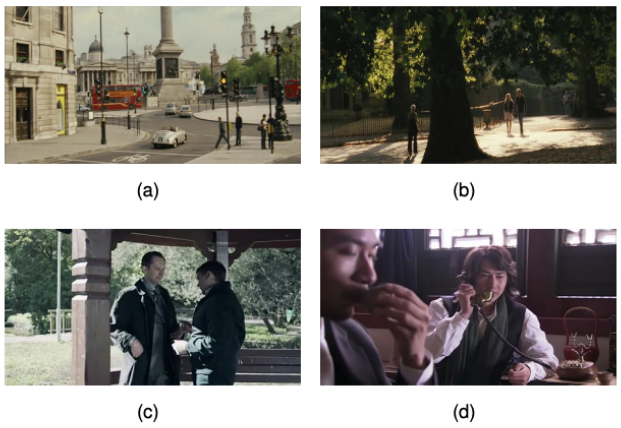

Figure 1. Different types of shot cuts, such as cutting on action (a), dissolve (b), cut-in (c), and cross-cut (d). **Clickable**: best viewed with Adobe Acrobat Reader; click to watch the animation.

Challenge

TV 쇼와 영화에서의 TAL(temporal action localization)의 가장 중요한 점은 short cuts에서 생기는 event의 분산입니다. 장면의 각도와 깊이는 많은 변화가 생깁니다.

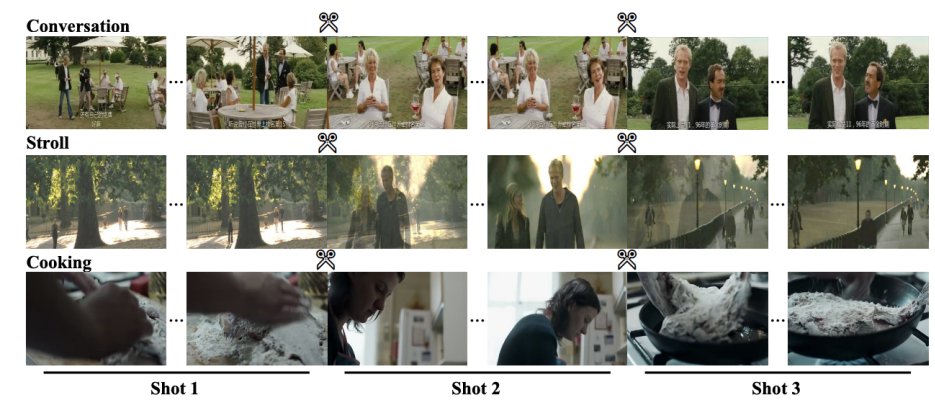

Figure 2. Examples of multi-shot events. In each row, we show three consecutive shots in an instance and select two frames per shot for illustration. The scissor icons indicate the shot boundaries.

shot cuts으로 인해 장면 전환, 배우 변경, 사이드 효과가 발생합니다. 이러한 한 이벤트 내에서의 다양성은 TAL을 더욱 어렵게 합니다.

Our Contribution

Multi-shot temporal event localization을 발전시키기 위해 우리는 유저들이 생성한 THUMOS14, ActivityNet-1.3 등 여러 데이터셋과 다른 비디오 편집 기술이 들어간 드라마 비디오로부터 만든 MUSES 데이터셋을 제시합니다. 176개의 비디오가 19개의 shots으로 구성되어 있습니다. 그리고 large scale dataset으로 31,377 event가 716시간의 비디오에서 존재합니다. 우리는 MUSES를 기존의 sota 모델인 P-GCN, G-TAD, MR로 평가를 해보았으나 P-GCN에서 가장 높은 점수인 13.1%mAP score가 나왔고 이는 MUSES dataset이 어려움을 입증합니다. 그리고 우리는 이러한 데이터셋에 알맞는 model baseline을 제시하며 이 baseline은 18.9%mAP score와 THUMOS14에서는 56.9%mAP score를 달성했습니다.

Related Work