

基于音视频融合的电影暴力内容检测研究

龚 语

导师：王伟强

摘 要

随着网络技术和电影工业的发展，越来越多的视频内容广泛流传，而其中或多或少地存在着一些暴力内容。一方面，暴力场面容易吸引观众的注意，属于影片中的精彩内容，研究和分析暴力内容对自动生成影片摘要和精彩内容检索具有辅助性研究价值；另一方面，通过对暴力程度的鉴定，可以过滤掉一些不适宜儿童观看的过度暴力的镜头，因此，合适的暴力评定方法有利于推动电影分级等相关工作的发展。人工标识与过滤暴力内容也终将因成本高、速度慢而跟不上视频产业发展的步伐，所以人们期待它可以被自动识别与标注技术所取代。

本文在分析了电影拍摄手法和表现手段的基础上，对影视内容中的暴力事件展开研究，并建立了层次型的暴力内容自动检测模型。首先利用半监督的学习方法，从连续的视频流中检测出快节奏、大音量、少对话和轻音乐的候选暴力镜头；继而在这些候选镜头中进一步检测爆炸、枪击、打斗、刀剑、撞击、引擎等具体的暴力相关事件，并推断它们与暴力的相关程度，从而根据其在镜头中的分布，计算出镜头的暴力分值；最终将这两个阶段的结果融合，对整部影片得到以镜头为单位的连续的暴力程度曲线。本文的主要贡献如下：

- 1) 提出了通用的暴力内容自动检测框架。该框架从暴力内容的特点出发，不仅能够检测由“枪击-爆炸-追车”等因素构成的好莱坞式暴力场景，也适用于以“搏击-刀剑”构成的冷兵器暴力场景检测。在音视频特征集合上，以镜头为基本单位，一方面把暴力理解为一种统一的抽象概念，检测出快节奏的镜头作为候选检测集，将暴力场景与慢节奏舒缓场景予以区分；另一方面用一些具体事件的组合来描述暴力，通过检测这些暴力相关事件，将暴力内容与体育运动等快节奏视频予以区分。
- 2) 改进了 SCFL 半监督学习方法用于候选暴力镜头检测。由于有监督学习需要大量的训练样本才能正确地描述特征空间分布，而样本标注工作耗时费力，需要投入大量人力。考虑到标注成本和特征集合自身的特点，本文在音视频底层、中层特征构成的正负视图基础上，改进了 SCFL 的数据选择功能，在扩充训练集的同时有效地控制了融入学习的数据质量，从而以半监督的方式高效地检测出候选暴力镜头。经实验证

明，该算法的正确率比有监督的 SVM 方法和传统的 SCFL 方法都有所提升。

- 3) 提出了适宜应用需求的暴力相关事件检测算法。电影中的声音条件非常复杂，尤其在暴力场景中，各种声音的交错重叠使音频关键字检测很难达到令人满意的准确率。因此，基于音频底层特征，本文提出了侧重于将暴力相关事件与其他事件分开的算法，而不盲目追求单一暴力事件的检测效果。不仅利用了概率输出的 SVM 与融合时间信息的 HMM 模型，还计算了各事件与暴力的相关系数，从而由具体事件推断出镜头的暴力分值。
- 4) 基于本文的层次模型，设计并实现了电影暴力内容检测系统。该系统可以随着电影的播放实时地检测镜头并显示当前的暴力程度值，赋予了用户直观的感官体验。

总之，本文的研究工作是基于用户迫切的应用需求和广泛的应用前景而展开的，通过对暴力镜头检测、暴力事件检测等技术的研究，为用户更好地监测影视文件中的暴力内容提供了解决方案。

关键词： 暴力内容检测，候选暴力镜头检测，半监督学习，暴力相关事件，语义推断

Detecting Violent Scenes in Movies by Auditory and Visual Cues

Gong Yu (Computer Applications)

Directed By Wang Weiqiang

With the development of Internet technology and film industry, more and more movies are widely spread, and some of them contain violent scenes more or less. On one hand, audiences are prone to be attracted by violent scenes, and they are usually considered as highlights of a film, so violence detection would be very useful for movie skimming and highlight extraction. On the other hand, most parents in the world generally do not hope their children exposed to scenes with too much violent content, so automatic detection of violent scenes are also useful techniques to help parents to prevent their children from watching such movies. Therefore, manual labeling is expected to be ultimately replaced by automatic recognition and indexing techniques, due to its high labor cost, and slow efficiency relative to the rapid pace of film industry.

Based on the film-making rules and presentation patterns, we investigate the techniques of detecting the violent content in movies automatically, and propose a hierarchical model. Firstly, candidate violent shots with fast tempo, high loudness, less speech, and less light music are detected from continuous video streams in a semi-supervised way. Secondly, typical violence-related audio effects, such as explosions, gunshots, struggle, weapons, smashes, engines, are further detected for the candidate shots, and we manage to transform the confidences outputted by the classifiers of various audio events into a shot-based violence score. Finally, the first two-stage probabilistic outputs are integrated in a boosting way to generate a final violent score for each shot of the movie. The four main contributions of the thesis are summarized as follows:

- 1) Proposed a unified framework on automatic violence detection. Based on the characteristics of violence, the framework could cope with not only the Hollywood violent scenes composed of explosions, gunshots and car-racings, but also those person-on-person violent contents with elements such as fighting, swords, etc. On the basis of shot detection, it identifies violence from two layers by audio-visual cues. At first, by taking violence as a universe concept, it detects fast-paced scenes as candidate violent content to distinguish violence from low-paced scenes. Furthermore, violence is depicted by some typical events generally associated with violence, by which fast-paced sports content could be

removed from the candidate set.

- 2) Modified the Semi-supervise Cross Feature Learning (SCFL) Method and applied it in candidate violent shot detection. The performance of supervised learning could be guaranteed if only large numbers of training samples are collected to describe the distribution of feature spaces correctly. While labeling is such a time-consuming work, a great amount of manual labors should be invested in. Considering the labeling cost and the characteristics of our feature set, we modifies a semi-supervised learning algorithm based on SCFL to locate candidate violence shots, which only leverages abundant unlabeled data to boost the classification accuracy, but also guarantees the quality of unlabeled data added to training sets. The experiment shows our modified SCFL performs better than the traditional supervised SVM method and the original SCFL.
- 3) Proposed an application-driven violent event detection algorithm. Since the audio effects always consist of complicated sound conditions in movies, especially in the violent scenes where multiple events occur simultaneously, it is difficult to reach a satisfying accuracy on audio event detections. Therefore, we propose a strategy which does not pursue a high accuracy of individual event detection, but focus on how to separate violence-related audio events from others. SVMs and HMMs are exploited to model the events. Additionally, we use an overall score obtained by summing over weighted confidences for each audio effect as a confidence measure of violence, so that the semantic of violence is inferred.
- 4) Designed and developed a movie violence detection software system based on the hierarchical framework we proposed. When a movie is playing, the system performs real-time shot detection and violence score computation of the current shot, presenting a visualized impression to users.

In a word, the topic of the thesis is driven by practical application requirements. The proposed methods provide a solution which can help users better manage their movie resources and identify the violent content in movies more effeciently.

Keywords: Violence Detection, Candidate Violent Shot Detection, Semi-supervised Learning, Violence-related Events, Semantics Inference