此有学者采用运动特征来提取关键帧, 如通过分析视频帧的光流场进而根据运动场的变化提取运动特征, 虽然相对颜色等特征, 这种方法提高了准确度但光流场特征的提取通常比较复杂. 本文选取并融合了图像的颜色特征和图像目标的形状特征作为传统手工特征.

在 1989 年 Yann LeCun 初次提出“卷积”的概念, 并构建应用于图像分类的卷积神经网络模型 LeNet. 在 ILSVRC-2012 比赛中, Krizhevsky 等人设计出深度卷积网络模型 AlexNet[1], 将图像分类错误率从 26.2% 降到了 15.3%, 识别准确度远高于其他方法, 这促进卷积神经网络在视觉图像领域得到快速的发展, 发展至今, 其在图像方面显示出了更优秀的表现. 因此本文使用卷积神经网络提取特征向量作为视频帧的深度特征, 然后选择合适的图像相似度度量方法计算图像间相似性.

基于以上思想, 本文主要有以下 3 个方面工作: (1) 相对以往固定阈值的方法, 本文采用自适应阈值, 动态获取视频的关键帧数量; (2) 分别提取深度特征与手工特征并计算相似度, 融合两者相似度提取关键帧; (3) 对比 3 种视频关键帧提取方法实验数据, 验证本文算法的有效性.

1 相关研究

早期对关键帧的提取大多是基于图像的底层特征, 主要包含图像颜色特征、图像纹理特征、图像形状特征等[2]. 对于颜色特征的提取方法通常利用 RGB 空间的颜色直方图、HSV 空间的颜色直方图、颜色聚合向量等[3]; 对纹理特征的提取方法通常利用 LBP 方法、马尔可夫随机场模型法、灰度共生矩阵等; 对形状特征的方法通常利用几何参数法、傅里叶形状描述法、小波描述子等. 现有的特征提取方法大部分都是基于一种或多种特征的融合, 但图像的底层特征通常提取有限, 无法获取图片高级特征, 虽然目前提取的效果不错, 但仍有待提高.

随着深度网络结果的发展, 人们发现对于视频类的图像分析, 用卷积神经网络通过二维卷积核对视频帧进行滑动卷积操作,如图 1 所示, 对视频帧底层特征进行抽象提取并组合, 最终可获得视频帧更深层次特征的抽象描述. 然而单个二维卷积核不能很好提取视频帧时间特性, 所以文献[4]提出 3D 卷积神经网络 (3D Convolutional Neural Networks), 如图 2 所示. 3D-CNN 对相邻的 3 张视频帧用 3 个二维卷积核卷积, 并将卷积的结果相加, 从而提取了某种时间的相关性, 因此对特征的描述更为充分.

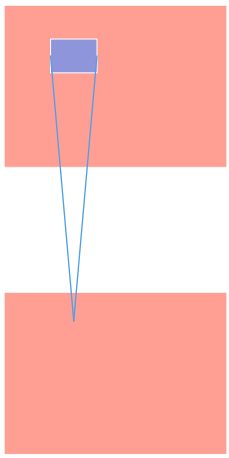

图 1 2D 卷积

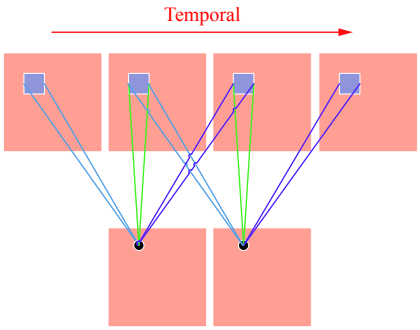

图 2 3D 卷积

2 基于融合特征的关键帧提取方法

本文的底层手工特征由将颜色直方图特征和方向梯度直方图表示, 深度特征通过 3D 卷积神经网络提取, 最后将深度特征向量相似度和手工特征向量相似度进行加权融合的方法进行相似度计算, 最后得到视频的关键帧. 整体结构流程如图 3 所示.

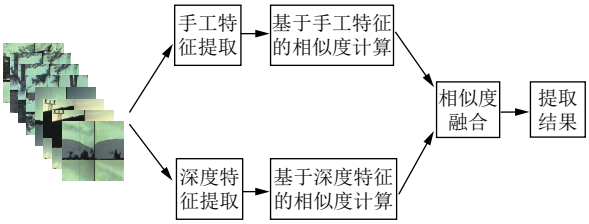

图 3 整体结构图

2.1 视频帧手工特征的提取

HSV (Hue, Saturation, Value) [5]颜色空间的概念是

Smith AR 于 1978 年初次提出的, 其中 H 表示色相, S 表示饱和度, V 表示明度. 色相 H 表示色彩属性, 范围区间[0°, 360°], 其中 0°表示红色, 120°表示绿色, 240°表示蓝色[6], 整体呈为环形, 色调随着角度的变化而变化. 饱和度 S 表示颜色的深浅, 取值区间为 0%~100%, 一般认为 S 值越高, 颜色就越深, S 取 0 时为灰度图像. 明度 V 表示色彩的明暗程度, 范围区间也是 0%~100%, 随 V 值的增大, 色彩逐渐变暗. HSV 颜色空间模型是 RGB 颜色空间的另一种表示方式, 但 HSV 颜色空间模型相对来说更为直观, 所以实际应用中更为广泛. 视频帧为 RGB 表示, 本文要从视频帧中提取颜色特征需要将视频帧转换为 HSV 表示, 如图 4 所示.

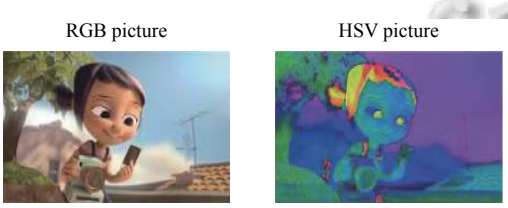

图 4 RGB2HSV 示意图

根据式 (1) 对 H、S、V 三通道特征量化构造特征矢量

$$G = HL_sL_v + L_vS + V \tag{1}$$

其中, L_s、L_V 分别为 S 通道和 V 通道的量化因子, 量化比例为 16:4:4. 通过 HSV 颜色空间的 3 个通道颜色特征, 可以得到每个通道上像素的分布, 从而获取到每个像素值对应的光谱信息, 将颜色空间进行颜色量化, 得到视频帧的量化颜色直方图, 如式 (2) 表示:

$$h(i,j,k) = \frac{N_{i,j,k}}{M \times N}(1 \le i \le 16, 1 \le j,k \le 4) \tag{2}$$

其中, $N_{i,j,k}$ 表示满足图像中 H 分量上第 i 个值、S 分量上第 j 个值以及 V 分量上第 k 个值的像素点个数, M 表示图像像素点总个数.

定义 $H_n(I)$ 和 $H_{n+1}(I)$ 分别是视频第 n 帧和第 n+1 帧图像的颜色直方图, 则两帧图像之间的相似度 S_{HSV} 可以用两者之间的余弦距离 $D(H_n, H_{n+1})$ 近似表示, 余弦计算公式如式 (3) 所示. 余弦距离范围是 0~1, 值越小, 则表示两帧图像越相似, 反之表示差异越大.

$$D(H_n, H_{n+1}) = \frac{H_n \cdot H_{n+1}}{|H_n| \times |H_{n+1}|} \tag{3}$$

颜色直方图不关心色彩所处的位置, 对视频帧中由于光照变化带来的阴影干扰, 抖动等有很好的区分去除能力, 同时颜色直方图对背景的干扰也有很好的抑制作用, 因此可以用来增强关键帧提取的抗噪性.

本文采用方向梯度直方图 HOG (Histogram of Oriented Gradient)[7]来表征视频帧的目标对象形状特征. 方向梯度直方图的重要思想是像素梯度或边缘的方向密度分布能够很好地表示图片中的目标形状. 对梯度直方图的计算首先对图像进行标准化处理, 之后用梯度算子[–1, 0, 1]及其转秩对视频帧分别进行卷积运算[8], 从而得到 x 方向和 y 方向的梯度分量 *xGradient* 与 *yGradient*. 最后分别用式 (4)、式 (5) 计算出像素点的梯度大小和方向.

$$G_x(x,y) = H(x+1,y) - H(x-1,y) \tag{4}$$

$$G_y(x,y) = H(x,y+1) - H(x,y-1) \tag{5}$$

式中, $H(x, y)$, $G_x(x, y)$, $G_y(x, y)$ 分别为输入的视频帧在像素点 (x, y) 处的像素值、水平方向梯度、垂直方向梯度[9]. 像素点 (x, y) 处的梯度幅值和梯度方向分用式 (6)、式 (7) 所示:

$$G(x,y) = \sqrt{G_x(x,y)^2 + G_y(x,y)^2} \tag{6}$$

$$\alpha(x,y) = \tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{7}$$

将视频帧进一步划分为若干单元块, 对单元块内若干 cell 中每个像素点根据梯度方向做统计分析, 得到以梯度方向为坐标轴的直方图[9], 然后对 cell 组成块并进行块内归一化, 归一化公式如式 (8) 所示. 将所有块的特征向量组合起来即可得到目标对象的特征向量.

$$u = \frac{v}{\sqrt{\|V\|_2^2 + \delta^2}} \tag{8}$$

式中, V 表示包含给定块的统计直方图信息的未归一化向量, δ 为趋于零的常数, $\|V\|_2$ 为 v 的 2-范数. 假定第 i 帧整体特征向量用 V_i 表示, 第 i+1 帧用 V_{i+1} 表示, 则两帧的相似度 S_{HOG} 可根据向量夹角余弦值表示, 值越接近 1 则方向越吻合, 两帧的相似度也越高, 余弦值的计算如式 (9) 所示.

$$\cos(\theta) = \frac{V_i \cdot V_{i+1}}{\|V_i\| \cdot \|V_{i+1}\|} = \frac{\sum_1^n (V_i \times V_{i+1})}{\sqrt{\sum_i^n (V_i)^2} \times \sqrt{\sum_{i+1}^n (V_{i+1})^2}} \tag{9}$$

2.2 视频帧深度特征的提取

3D-CNN 结构由一个硬连接线层、3 个卷积层、2 个下采样层, 1 个全连接层组成[4]. 本文提出用 3D-CNN 来提取视频帧的深度特征, 计算其相似度, 并与传统手工提取特征计算的相似度进行加权融合, 进而根据融合相似度提取出视频的关键帧. 对于深度特征, 首先取视频中连续帧作为 3D-CNN 的输入, 经过第一层硬连线 (hardwired) 层编码获得视频帧的灰度、梯度以及光流特征信息, 其中梯度描述视频帧的边缘分布, 光流描述目标的运动趋向, 然后将梯度信息和光流信息作为下一层卷积层的输入进行后续识别处理. 在像素值 (x, y) 处, 提取的特征单位值用 V_{ij}^{xyz} 表示, i 表示层数, j 表示特征图序号, 单位值计算方法如式 (10) 所示.

$$v_{ij}^{xyz}=\tanh\left(b_{ij}+\sum_{n}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}w_{ijn}^{pqr}v_{(i-1)n}^{(x+p)(y+q)(z+r)}\right) \quad (10)$$

其中, b_{ij} 表示特征图的偏置值, W_{ijn}^{pqr} 是连接第 n 个特征图的核第 (p, q, r) 的值, P_i 和 Q_i 表示核的高和宽, R_i 表示卷积核在时间维度的大小.

通过多次卷积核卷积和下采样后, 3D-CNN 将输入的连续视频帧转换为特征向量表示, 这里, 我们去掉原网络结构中最后的全连接层, 选择最后一个卷积层的 feature map 作为要提取的 n 帧特征向量 G_{df}.

由于深度卷积采样到的特征具有高维度的特性, 因此本文使用哈希 (Hashing) 算法处理图像的深度特征. 哈希算法将高维数据编码为一组二进制代码, 并能维持图像或视频高维数据的元相似性[10]. 本文在卷积层后加入了 Hash 层, 用 Sigmoid 函数作为卷积网络的激活函数[11], 将特征值限制在 0~1 之间, 构造特征的 Hash 码, 最后通过 Hash 码计算视频帧的 Hamming 距离, Hamming 距离越小则表示视频帧的相似度 S_{df} 就越高. 假设两帧的 Hash 码分别为 α、β, 则 Hamming 距离 D 定义如式 (11) 所示.

$$\begin{cases}\alpha=\alpha_1,\alpha_2,\cdots,\alpha_n\\ \beta=\beta_1,\beta_2,\cdots,\beta_n\\ I(\alpha_i,\beta_i)=\begin{cases}1, & \text{if} \quad \alpha_i\neq\beta_i\\ 0, & \text{if} \quad \alpha_i=\beta_i\end{cases}\\ D=\sum_{1}^{n}I(\alpha_i,\beta_i)\end{cases} \quad (11)$$

2.3 基于深度特征与手工特征融合的关键帧提取

基于传统手工特征和深度特征的关键帧提取分为两步, 首先使用传统手工方法提取出视频帧的手工特征, 然后用 3D-CNN 提取视频的深度特征, 由于两者特征维度的不同, 所以分别计算两者的相似度. 首先根据 2.1 节计算手工特征颜色直方图和方向梯度直方图特征的余弦距离得到传统手工特征的相似度 S_{HSV} 和 S_{HOG}, 然后根据 2.2 节通过哈希算法计算得到深度特征的哈希码, 并通过 Hamming 距离得到深度特征的相似度 S_{df}, 最后融合两种特征的相似度作为提取视频关键帧的依据.

特征融合方法分为拼接融合、加权融合、基于系数特征表示理论的特征融合、基于贝叶斯理论融合等. 由于手工特征和深度特征有维度差异, 本文选择加权融合方式, 将两者相似度进行融合. 首先对两者相似度根据权重大小做加权处理, 然后线性融合传统特征和深度特征相似度, 避免了手工特征与深度特征的维度差异, 最后通过融合后的相似度根据阈值提取关键帧. 相似度 S 计算方法如式 (12) 所示:

$$S=\alpha\cdot S_{\mathrm{HSV}}+\mu\cdot S_{\mathrm{HOG}}+\beta\cdot S_{df} \quad (12)$$

式中, α、μ、β 分别为手工特征和深度特征的权重因子, 比例采用 1∶1∶2. 在相似度计算时为了使关键帧的数目根据视频内容自动调整阈值, 本文使用自适应阈值的方法设置相似度的阈值.

$$\varepsilon=\frac{1}{n}\sum_{i=1}^{n}S(f_{i+1},f_i)+\tau \quad (13)$$

式 (13) 中, ε 为相似度阈值, n 为总的视频帧数量, f_i 表示当前帧, τ 为域值的自适应调节因子. 本文总体算法步骤如下所示:

```
Begin
将视频分割为视频帧集 F {f1,f2,f3,⋯,fn};
定义空的关键帧集合 KF{};
输入融合后的视频级相似度集 S{s1, s2, ⋯, sn};
For i=1: n;
If (相似度 S>阈值 ε)
    Then 将 fi+1 放入关键帧集 KF{}
Else   i++;
  End if
End for i
输出采集到的视频关键帧集合 KF{kf1, kf2, ⋯}
End
```

3 实验及分析

在本节中, 为验证本文算法的有效性, 本文使用 *Xshell* 远程工具在服务器上搭建 *PyTorch* 深度学习框

架, 使用 *python3.6* 进行实验及其相关分析. 为了度量不同方法的实验结果, 本文分别使用查准率、查全率、F_1 度量来评估算法的性能[3], 公式如式 (14) 所示.

$$\begin{cases} 查准率(precision) = \dfrac{TP}{TP+FP} \\ 查全率(recall) = \dfrac{TP}{TP+FN} \\ F_1 = \dfrac{2\times TP}{sum+TP-TN} \end{cases} \tag{14}$$

其中, *TP* 表示真正例, *FN* 表示假反例, *FP* 表示假正例, *TN* 表示真反例, F_1 是基于查准率和查全率的调和平均分数.

本文实验视频集从公开视频项目 Open Video Project[12]网站上下载得到, 下载的视频集共分为 5 类, 其中记录片、教育、历史、公共服务各选 4 个视频, 并随机从 Youtube 网站另外选择 4 个视频, 共 20 个视频构成实验数据集. 为验证算法的有效性, 本文选择两种常用方法进行对比实验, 一种是基于帧间差分[13]的方法, 一种是基于感知哈希算法[14]的方法. 实验从 5 类视频集中各选择一个代表视频进行实验, 3 种算法提取的结果统计情况如表 1 所示, 其中 Video3 的可视化效果如图 5–图 7 所示.

图 5 基于帧差局部最大值提取结果

图 6 基于感知 Hash 匹配提取结果

图 7 本文算法提取结果

Video3 是从长历史片中截取的一段, 描述了生态学家研究云对麋鹿觅食的影响. 从图 5–图 7 可以看出基于感知哈希匹配相似度的方法提取效果最差, 不仅存在冗余帧, 而且存在大量漏检帧, 基于帧差法的提取结果与本文结果数量相似, 但本文提取的结果比帧差法提取结果更丰富, 漏检帧更少.

表 1 中 *A* 表示基于帧差法提取算法, *B* 表示基于感知 Hash 相似度匹配算法, *C* 表示本文算法. 由表中数据可以看出 3 种算法中, 基于感知 Hash 匹配相似度的算法 F_1 值普遍偏小, 基于帧差法的 F_1 值与本文算法得到的 F_1 值相比, 本文算法在 Video5 视频类型上与帧差提取算法有一定差距, 这是因为 Video5 视频整体色彩变化不明显, 所以本文的手工提取特征部分提取效果稍差. 但从整体来看, 本文算法比帧差法和感知 Hash 匹配法提取效果更好, 准确率更高, 冗余度更小, 提取结果可以更全面的描述视频内容.

表 1 对比实验统计结果

Video	总帧数 (帧)	真关键帧数 (帧)	提取的关键帧 (帧)			查准率 (%)			查全率 (%)			F_1		
			A	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
Video1	1813	22	27	16	25	74	88	84	91	63	95	81	73	89
Video2	1798	29	30	10	32	70	90	90	72	31	93	71	46	91
Video3	1186	15	17	8	19	71	75	63	73	40	80	72	52	70
Video4	1905	28	30	22	30	77	82	83	82	64	89	79	72	86
Video5	3555	42	62	18	54	86	94	67	90	40	86	88	56	75

4 结束语

本文提出基于融合特征的视频关键帧提取的方法, 充分利用了传统手工特征和深度特征的特点及优势. 将提取到的视频图像的传统手工特征与基于深度神经网络提取的深度特征计算得到相似度并进行融合, 以自适应阈值作为门限提取关键帧. 通过对公共视频集进行实验, 实验结果表明对关键帧提取有更为准确和全面的提高, 与传统方式提取的方法相比, 本文方法提取的特征更丰富, 提高了视频关键帧的准确度并在冗余度方面也有良好的表现, 对视频的分析研究具有重要的作用.

参考文献

1 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

2 韩震, 刘志远. 基于多特征融合和二维投影非负矩阵分解

的图像检索. 价值工程, 2016, 35(8): 228–230, 231. [doi: 10.14018/j.cnki.cn13-1085/n.2016.08.087]
3 王金娟. 基于颜色特征的图像检索技术研究[硕士学位论文]. 长沙: 湖南大学, 2010.
4 Ji SW, Xu W, Yang M, *et al*. 3D Convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/tpami.2012.59]
5 Smith AR. Color gamut transform pairs. Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques. New York, NY, USA. 1978. 12–19. [doi: 10.1145/800248.807361]
6 赵正利. 一种基于高斯混合模型的图像检索方法[硕士学位论文]. 青岛: 中国海洋大学, 2007.
7 Dalal N, Triggs B. Histograms of oriented gradients for human detection. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, Chile. 2005. 886–893. [doi: 10.1109/cvpr.2005.177]
8 朱志磊. 关于 ATM 机隔间内尾随检测算法的研究[硕士学位论文]. 杭州: 浙江工业大学, 2015.
9 王雪锋, 叶飞. 一种基于 HOG 与 SVM 的监控视频车辆识别方法. 微型机与应用, 2013, 32(17): 34–37. [doi: 10.3969/j.issn.1674-7720.2013.17.013]
10 王小凤, 张飞, 耿国华, 等. 一个基于深度图像的三维模型检索算法. 计算机工程与应用, 2012, 48(7): 197–200. [doi: 10.3778/j.issn.1002-8331.2012.07.053]
11 李蕾. 基于哈希的图像检索研究[硕士学位论文]. 北京: 北京交通大学, 2017.
12 The Open Video Project. https://open-video.org/.
13 陈宝远, 霍智超, 陈光毅, 等. 一种改进的三帧差分运动目标检测算法. 应用科技, 2016, 43(2): 10–13. [doi: 10.11991/yykj.201506027]
14 Zauner C, Steinebach M, Hermann E. Rihamark: Perceptual image hash benchmarking. Proceedings of SPIE 7880, Media Watermarking, Security, and Forensics III. San Francisco, CA, USA. 2011. 7880. [doi: 10.1117/12.876617]