

视频质量评估方法、装置、设备及存储介质

本申请要求在2018年12月28日提交中国专利局、申请号为201811627024.4的中国专利申请的优先权，该申请的全部内容通过引用结合在本申请中。

技术领域

本申请实施例涉及视频检测技术领域，例如涉及一种视频质量评估方法、装置、设备及存储介质。

背景技术

随着互联网技术的发展，短视频类应用程序（Application，APP）的使用越来越广泛。短视频类APP每天获取用户制作并上传的大量短视频，这些短视频在剔除了违规视频以后，需要从剩下的视频中挑选出优质视频，向观众推送。优质视频是指潜在的会受观众喜欢的视频。能否准确且全面的挑选出剩下的视频中的优质视频来展现给观众，决定了短视频类APP的观众观看时长、保存量、下载量等关键指标。

相关技术中，常用的优质视频评估方法主要有两类。一类是根据视频的外部信息来评估视频是否优质，也就是在视频发布出来后，根据观众的反馈数据来评估视频的优质程度。这种方法的缺点在于需要等视频发布并收集到观众反馈后才能对视频的优质程度进行评估，缺乏时效性。另一类是根据视频内部的一些简单的图像参数来评估视频内容是否优质。例如用视频的封面或关键帧进行图片的分类或聚类，形成对视频内容的一些简单的描述信息，根据描述信息判断视频内容是否优质。此方法的缺点在于使用的图像参数太简单，无法准确和全面的描述视频的内容，进而无法对视频优质程度进行准确评估。

发明内容

本申请提供一种视频质量评估方法、装置、设备及存储介质，提高了视频质量评估的时效性，提升了优质视频评估的准确度。

本申请实施例提供了一种视频质量评估方法，包括：

对待评估视频进行解码，得到所述待评估视频的图像序列和音频信息；

提取所述图像序列中的动作特征向量和人脸特征向量，以及提取所述音频信息中的音频特征向量；

根据所述动作特征向量、所述人脸特征向量和所述音频特征向量构建视频

特征向量;

根据所述视频特征向量确定所述待评估视频的质量评分。

本申请实施例还提供了一种视频质量评估装置,包括:

视频解码模块,设置为对视频进行解码,得到所述待评估视频的图像序列和音频信息;

向量提取模块,设置为提取所述图像序列中的动作特征向量和人脸特征向量,以及提取所述音频信息中的音频特征向量;

向量构建模块,设置为根据所述动作特征向量、所述人脸特征向量和所述音频特征向量构建视频特征向量;

质量评分模块,设置为根据所述视频特征向量确定所述待评估视频的质量评分。

本申请实施例还提供了一种设备,所述设备包括:

一个或多个处理器;

存储器,设置为存储一个或多个程序;

当所述一个或多个程序被所述一个或多个处理器执行,使得所述一个或多个处理器实现本申请任意实施例提供的视频质量评估方法。

本申请实施例还提供了一种计算机可读存储介质,其上存储有计算机程序,该程序被处理装置执行时实现本申请任意实施例提供的视频质量评估方法。

附图说明

图1是本申请实施例一提供的一种视频质量评估方法的流程图;

图2是本申请实施例一提供的一种视频质量评估方法应用场景的示意图;

图3是本申请实施例二提供的一种视频质量评估方法的流程图;

图4是本申请实施例三提供的一种视频质量评估方法的流程图;

图5是本申请实施例四提供的一种视频质量评估装置的结构示意图;

图6为本申请实施例五提供的一种设备的结构示意图。

具体实施方式

下面结合附图和实施例对本申请进行说明。本文所描述的具体实施例仅仅用于解释本申请,而非对本申请的限定。为了便于描述,附图中仅示出了与本申请相关的部分而非全部结构。

实施例一

图 1 是本申请实施例一提供的一种视频质量评估方法的流程图，本实施例可适用于评价视频质量的情况，该方法可以由视频质量评估装置来执行，该视频质量评估装置可以由硬件和/或软件实现。该视频质量评估装置可以由多个物理实体构成，也可以是由一个物理实体构成。一实施例中，该该视频质量评估装置集成在计算机设备中。

如图 1 所示，本实施例提供的视频质量评估方法，包括如下步骤：

S110、对待评估视频进行解码，得到待评估视频的图像序列和音频信息。

本实施例中，视频类服务器每天会接收到大量由用户制作并上传的视频。视频类服务器需要从用户上传的所有视频中剔除掉违规视频，并筛选出观众可能喜欢的视频。因此，需要对用户上传的视频进行检测、筛选，并根据筛选结果将观众可能喜欢的视频推荐给观众。

本实施例提供的视频质量评估方法可以由视频质量评估装置来执行。图 2 是本申请实施例一提供的一种视频质量评估方法应用场景的示意图。如图 2 所示，用户通过用户端 210 录制视频并上传至服务器 220 之后，服务器 220 中的后台 221 或者推荐系统 222 调用视频质量评估装置 223，视频质量评估装置 223 对用户上传的视频进行评估并输出一个评估结果，将评估结果返回至后台 221 或者推荐系统 222，后台 221 或者推荐系统 222 根据评估结果对该视频进行处理。一实施例中，若评估结果是优质视频，则推荐系统 222 将该视频推荐给观众端 230。一实施例中，用户端 210 和观众端 230 可以是同一个终端设备，也可以是两个不同的终端设备。

本实施例中，待评估视频就是指用户上传至视频类服务器，视频质量还未知的视频。用户在上传视频时，为了数据传输的需求，需要对视频进行编码。编码是指通过特定的压缩技术，将一个格式的视频文件转换成另一种格式的视频文件。一实施例中，待评估视频至少包括图像和音频。待评估视频可以认为是由多张图像组成的集合。在对待评估视频进行评估之前，需要对已经编码的视频进行解码，视频可以认为是由多张快速播放的图像组成，由于肉眼的视觉暂留现象，才使人感觉画面是连续的。因此，将视频进行解码，可以得到多帧图像数据，每帧图像数据包括一张图像。将视频解码之后，可以得到由多张图像组成按一定顺序排列的图像序列。

本实施例中，对待评估视频进行解码后，还可以得到待评估视频的音频信息。音频信息包括待评估视频中的背景音乐和人物的语言信息。

S120、提取图像序列中的动作特征向量和人脸特征向量，以及提取音频信

息中的音频特征向量。

在本实施例中，动作特征向量是指以向量的形式表示的动作特征，动作特征至少包括待评估视频内的人物动作和周围环境。示例性的，人物动作是指人物的舞蹈动作、运动动作等。周围环境是指建筑物、高山、灌木丛等。动作特征很大程度上决定了视频的受欢迎程度。

在本实施例中，将图像序列中的每一帧图像的姿势向量提取出来，一帧图像对应一个矩阵信息，通过将该矩阵信息进行合并处理，可以得到对应的姿势向量。姿势向量反映的是每一帧图像所对应的图像信息，一个图像序列可以包含多帧图像，每一帧图像对应一个姿势向量。将得到的多个姿势向量进行聚类处理，将多个姿势向量转换成动作特征向量。本实施例中仅对动作特征向量的提取方法进行说明，而非限定。可以根据需要采用任意的动作特征向量的提取方法。

一实施例中，可以预先建立动作特征提取网络，利用动作特征提取网络对图像序列中的每一帧图像进行特征提取，得到图像序列中的动作特征向量。一实施例中，对动作识别网络进行微调，得到动作特征提取网络，将图像序列输入动作特征提取网络，利用动作特征提取网络对每一帧图像进行特征提取，得到图像序列中的动作特征向量。动作识别网络可以由动作分类数据集中取得较好效果的至少一类网络来组成，包括但不限于基于三维（three dimensions, 3D）卷积网络的 C3D，交互式三维（Interactive three dimensions, I3D）卷积网络等、基于光流和空间的双流（Two-Stream, 2-Stream）方法的时域分段网络（Temporal Segment Networks, TSN），以及基于循环神经网络（Recurrent Neural Network, RNN）的多种动作识别网络。

人脸特征向量是指用于描述人脸面部特征的向量。人脸特征向量的提取是指在人脸检测的基础上，在人脸所在的区域获取人脸面部特征向量的过程。

在本实施例中，人脸特征向量提取主要有两个步骤，首先对图像序列中的人脸进行检测，然后对检测到的人脸信息进行特征提取。一实施例中，首先，利用多任务卷积神经网络（Multi-task convolutional neural networks, MTCNN）方法对获取的图像序列进行人脸检测，得到人脸检测数据。然后，利用 FaceNet 方法对检测得到的人脸据进行特征提取，得到人脸特征向量。本实施例中，卷积神经网络是深度学习的一项主要技术，MTCNN 方法是用于人脸检测的一种方法，人脸检测是指判断一幅图像中是否出现人脸。FaceNet 是一种方法的名称，这个方法使用卷积神经网络，把人脸映射到欧几里得特征空间，得到人脸特征向量。

视频中的背景音乐或者人物语音也是视频的重要内容之一，需要提取出背

景音乐或者人物语音的音频信息的特征，根据音频信息的特征对视频的优质程度作一个预估。从视频中提取出音频信号后，需要先用一些传统的音频信号处理方法对音频进行预处理，提取出音频的频谱或倒谱特征，然后送入音频特征提取网络，提取出可用于音乐识别或者人物语音识别的音频特征向量。

音频特征向量是指描述视频中声音信息的向量。一实施例中，提取音频信息中的音频特征向量，是指提取音频信息中反应音频信息本质特征的成分，如：音频的中心频率，一些频率段内音频的能量特征或多个时间内音频的能量分布特征，并将这样特征组合起来，得到相应的音频特征向量。

S130、根据动作特征向量、人脸特征向量和音频特征向量构建视频特征向量。

在本实施例中，所述视频特征向量是指用于表征视频质量的向量。视频特征向量是根据动作特征向量、人脸特征向量和音频特征向量构建的特征向量。

将上述 S120 中提取的动作特征向量、人脸特征向量和音频特征向量进行合并，得到一个视频特征向量。一实施例中，若在待评估视频中仅仅提取到一种特征向量，那么则将未提取到的特征向量全部置 0，构成视频特征向量。示例性的，若在待评估视频中仅仅提取到动作特征向量，那么则将人脸特征向量和音频特征向量全部置 0，构成视频特征向量。若在待评估视频中仅仅提取到两种特征向量，那么则将未提取到的特征向量全部置 0，构成视频特征向量。示例性的，若在待评估视频中仅仅提取到动作特征向量和人脸特征向量，那么则将音频特征向量全部置 0，构成视频特征向量。

S140、根据视频特征向量确定待评估视频的质量评分。

在本实施例中，预先搭建和训练视频评估网络，将视频特征向量输入视频评估网络，视频评估网络输出一个评分作为待评估视频的质量评分。一实施例中，视频评估网络可以是一个简单的多层感知器（Multilayer Perceptron, MLP）也可以是较为复杂的 RNN 网络。一实施例中，建立优质视频训练数据集，根据上述 S120 中的方式对优质视频训练数据集提取相应的特征，并存储作为视频评估网络的训练数据。

一实施例中，在视频评估网络训练时，需要考虑到一些视频中没有人脸的情况，因此在训练时需要在优质视频集合中随机的将人脸特征向量替换为全 0，作为一种数据增强手段，使训练出来的网络能更好的适应实际的数据。

本实施例提供的视频质量评估方法，首先通过对待评估视频进行解码，得到待评估视频的图像序列和音频信息，然后提取图像序列中的动作特征向量和人脸特征向量，以及提取音频信息中的音频特征向量，再根据动作特征向量、

人脸特征向量和音频特征向量构建视频特征向量；最后根据视频特征向量确定待评估视频的质量评分。可见，本实施例可以根据待评估视频中提取的动作特征向量、人脸特征向量和音频特征向量中的至少一个向量，确定待评估视频的质量评分，提高了视频质量评估的时效性，提升了优质视频评估的准确度。

实施例二

图 3 是本申请实施例二提供的一种视频质量评估方法的流程图；本实施例可适用于评价视频质量的情况，本实施例对上述实施例中的视频质量评估方法进行说明，如图 3 所示，本实施例提供的视频质量评估方法，包括如下步骤：

S310、对待评估视频进行解码，得到待评估视频的图像序列和音频信息。

S320、利用动作特征提取网络对图像序列中的每一帧图像进行特征提取，得到图像序列中的动作特征向量。

在本实施例中，在利用动作特征提取网络对每一帧图像进行特征提取之前，对动作特征提取网络进行搭建和训练。一实施例中，搭建和训练动作特征提取网络主要包括如下步骤：获取训练视频集合，并获取训练视频集合中的每个训练视频中的图像序列；在视频数据库中训练出动作分类网络；使用所述训练视频集合中的每个训练视频中的图像序列对所述动作分类网络进行训练，并删除所述动作分类网络中的分类层，得到动作特征提取网络。

一实施例中，训练视频集合是指通过人工筛选的方法选出的优质视频集合。优质视频的种类和数量越多，动作特征提取网络的准确度越高，视频评估结果更加有效。首先，获取优质视频集合，将优质视频集合中的每个优质视频进行解码，得到每个优质视频对应的图像序列。

根据神经网络在视频数据库中训练出动作分类网络。所述神经网络包括但不限于基于 3D 卷积网络的 C3D, I3D 卷积网络等、基于光流和空间的 2-Stream 方法的 TSN，以及基于 RNN 的多种动作识别网络。所述视频数据库中包括但不限于 Kinetics, Youtube-8M 数据集。Kinetics 是一种大型公开数据集的名称，是一个具有精确标注动作的数据集。Youtube-8M 也是一种大型公开数据集的名称，是一个大型视频数据集。可以用 Kinetics 数据集或 Youtube-8M 数据集中视频进行训练，得到动作分类网络。在本实施例中，不对动作分类网络的训练方法进行限定，可以采用相关技术中任意的模型训练方法，训练得到动作分类网络。

得到动作分类网络之后，使用优质视频集合中的每个优质视频中提取的图像序列对动作分类网络进行微调，并删除动作分类网络中的分类层，得到动作特征提取网络。

S330、利用人脸特征提取网络提取图像序列中的人脸特征向量。

在本实施例中，人脸特征提取网络包括人脸检测子网络和特征提取子网络。一实施例中，利用人脸特征提取网络提取所述图像序列中的人脸特征向量包括：利用人脸检测子网络检测出图像序列中每一帧图像中的人脸图像；利用特征提取子网络对图像序列中每一帧图像中的人脸图像进行特征提取，得到图像序列中的人脸特征向量。

在本实施例中，人脸检测子网络可以由常见的人脸检测网络来构建，例如：MTCNN。一实施例中，利用人脸检测子网络检测出图像序列中每帧图像中的人脸图像包括：先利用 MTCNN 网络中的建议网络 P-Net 筛选出每帧图像中的候选人脸区域及对应的候选人脸区域边框，再利用极大值抑制算法对高度重叠的候选人脸区域边框进行合并；然后利用 MTCNN 算法中的完善网络 R-Net 对候选人脸区域进行细化，得到候选人脸图像，同时利用边界框回归算法和非极大值抑制算法删除错误的候选人脸区域边框；利用 MTCNN 算法中的输出网络 O-Net 过滤余下的候选人脸区域边框，将人脸从候选人脸图像中检测出来。一实施例中，当视频中存在多个人脸时，取较大且靠近图像正面的人脸。

一实施例中，人脸检测子网络可以直接使用公开的预训练好的人脸检测网络，也可以在优质视频集合上重新训练一个人脸检测网络，使人脸检测子网络在对优质视频识别人物上具有更高的准备度。

特征提取子网络可以由公开的预训练好的人脸识别网络构成，包括但不限于 facenet, insightface 等，也可以参考特征提取子网络的原理自行设计网络，然后在优质视频数据集合上面训练出专用的人脸识别模型。特征提取子网络训练过程中需要的同一个人的多张脸部图像可以从同一用户上传的自拍视频中获得。训练好后的特征提取子网络对同一个人的不同脸部图像会输出高度相似的特征向量，这个特征向量里包含了人的脸部相貌信息，可以用来识别出不同的人，也可以作为对长相的受欢迎程度进行预测的依据。

本实施例提供一种特征提取子网络训练方法。如下：获取训练视频集合，并提取训练视频集合中每个训练视频中的人脸图像；在人脸数据库中训练出人脸识别网络；使用训练视频集合中每个训练视频中的人脸图像对所述人脸识别网络进行训练，得到特征提取子网络。训练视频集合是指通过人工筛选的方法选出的优质视频集合。

S340、提取音频信息的倒谱系数和/或频域，利用音频特征提取网络对倒谱系数和/或频域进行特征提取，得到音频特征向量。

视频中的音乐或者人物语音也是视频的重要内容之一，需要使用音频特征

提取网络来提取出音乐或者人物语音的声音的特征向量，从而根据声音的特征向量对视频的优质程度作一个预估。

在本实施例中，倒谱系数为梅尔频率倒谱系数（melfrequency cepstrum，MFCC）。梅尔频率是基于人耳听觉特性提出来的，梅尔频率与赫兹（Hz）频率成非线性对应关系。MFCC 系数则是利用梅尔频率与 Hz 频率的这种关系计算得到频率特征。本实施例中，不对提取音频信息的倒谱系数的方法进行限定，可以选用任意的方法提取音频信息的倒谱系数。

一实施例中，频谱是指在音频信息进行分析时，分析音频信息中和频率有关部份，而不是和时间有关的部分。提取音频信息的频谱是指提取音频信息中和频率有关的部分。本实施例中，不对提取音频信息的频域的方法进行限定，可以选用任意的方法提取音频信息的频域。本实施例中可以提取音频信息的倒谱系数或频域中的任意一种，或者同时提取音频信息的倒谱系数和频域。

一实施例中，提取音频信息的倒谱系数和/或频域之前还包括：从视频中提取出音频信号后，需要先用一些传统的音频信号处理方法对音频进行预处理。一实施例中，预处理至少包括降噪、增益和数据归一化等处理。对音频信息进行预处理可以提高音频识别的正确率。音频的预处理使用的算法需要根据预训练的数据集的要求来决定，在训练和预测时保持一致。

本实施例中，提取音频信息的倒谱系数和/或频谱之后，利用音频特征提取网络对倒谱系数和/或频谱进行特征提取，得到音频特征向量。

本实施例中，提供一种音频特征提取网络的训练方法，获取训练视频集合，并提取训练视频集合中的每个训练视频中的音频信息；在音频数据库中训练出音频识别网络；使用训练视频集合中的每个训练视频中的音频信息对音频识别网络进行训练，得到音频特征提取网络。一实施例中，音频特征提取网络可以是任何基于卷积神经网络（Convolutional Neural Networks，CNN）或者 RNN 的网络，预先在公开的用于音乐识别或人物语音识别的数据集上训练音频识别网络，再在优质视频训练集合上细调音频识别网络，得到音频特征提取网络。

S320、S330 和 S340 三个步骤的执行顺序可以任意互换，本实施例中不对这三个步骤的执行顺序进行限定。

S350、根据动作特征向量、人脸特征向量和音频特征向量构建视频特征向量。

S360、将视频特征向量输入到视频评估网络中进行评估，得到待评估视频的质量评分。

在本实施例中，视频评估网络是可以是一个简单的多层感知器也可以是较

为复杂的 RNN 网络。视频评估网络的输入是视频特征向量，输出是一个对视频优质程度的评分。一实施例中，若输出评分为大于 0 分，则按实际输出评分作为待评估视频的质量评分。若输出评分为小于 0 分，则将 0 分作为待评估视频的质量评分。

一实施例中，所述视频评估网络包括第一层评估网络、第二层评估网络和输出层。

将所述视频特征向量输入到视频评估网络中进行评估，得到所述待评估视频的质量评分包括：将所述视频特征向量输入第一层评估网络，第一层评估网络通过递归和非线性运算获得初级特征向量；将所述初级特征向量输入第二层评估网络，第二层评估网络通过递归和非线性运算获得高级特征向量；输出层对初级特征向量和高级特征向量进行运算，得到所述待评估视频的质量评分。一实施例中，训练视频评估网络的时候，需要考虑到一些视频中没有人脸的情况，因此在训练时需要在优质视频集合中随机的将人脸特征替换为全 0，作为一种数据增强手段，使训练出来的模型能更好的适应实际的数据。

在上述实施例的基础上，可以将 S320 替换为提取图像序列中每一帧图像的光流数据；利用动作特征提取网络对图像序列中每一帧图像的光流数据进行特征提取，得到图像序列中的动作特征向量。

本实施例提供的视频质量评估方法，首先通过对待评估视频进行解码，得到待评估视频的图像序列和音频信息，然后利用动作特征提取网络对图像序列中每帧图像进行特征提取，得到图像序列中的动作特征向量，利用人脸特征提取网络提取所述图像序列中的人脸特征向量，利用音频特征提取网络对音频信息的倒谱系数和/或频域进行特征提取，得到音频特征向量，再根据动作特征向量、人脸特征向量和音频特征向量构建视频特征向量；最后根据视频特征向量确定待评估视频的质量评分。可见，本实施例可以根据待评估视频中提取的动作特征向量、人脸特征向量和音频特征向量中的至少一个向量，确定待评估视频的质量评分，提高了视频质量评估的时效性，提升了优质视频评估的准确度。

实施例三

在上述实施例的基础上，本实施例提供一种实例。图 4 是本申请实施例三提供的一种视频质量评估方法的流程图；如图 4 所示，上述视频质量评估方法包括如下步骤：

S4010、获取待评估视频。待评估视频可以是用户刚刚录制并上传至服务器的视频，也可以是服务器中存储的未进行质量评估的视频。一实施例中，本实

实施例中的待评估视频为用户录制的短视频。

S4020、将待评估视频进行解码，可以得到由多张图像组成按一定顺序排列的图像序列。

S4030、对待评估视频进行解码后，可以得到待评估视频的音频信息。音频信息包括待评估视频中的背景音乐和人物的语言信息。

S4020 和 S4030 的执行顺序可以互换，可以先执行 S4020 后执行 S4030，也可以先执行 S4030 后执行 S4020。一实施例中，还可以将 S4020 和 S4030 同时执行。本实施例中不对 S4020 和 S4030 的执行顺序进行限定。

S4040、提取图像序列中每一帧图像的光流数据，得到光流序列。

S4050、将 S4020 中解码得到的图像序列或 S4040 中得到的光流序列输入至动作特征提取网络。

S4060、动作特征提取网络输出一个动作特征向量。动作特征向量是指以向量的形式表示的动作特征的向量，动作特征至少包括待评估视频内的人物动作和周围环境。示例性的，人物动作是指人物的舞蹈动作、运动动作等。周围环境是指建筑物、高山、灌木丛等。动作特征很大程度上决定了视频的受欢迎程度。

S4070、将 S4020 中提取得到的图像序列，利用人脸检测子网络进行人脸检测，得到人脸图像。人脸检测子网络可以直接使用公开的预训练好的人脸检测网络，也可以在优质视频集合上重新训练一个人脸检测网络，使人脸检测子网在对优质视频识别人物上具有更高的准备度。

S4080、将 S4070 中得到的人脸图像输入至特征提取子网络。特征提取子网络可以由公开的预训练好的人脸识别网络构成，包括但不限于 facenet, insightface 等，也可以参考特征提取子网络的原理自行设计网络，然后在优质视频数据集合上面训练出专用的人脸识别模型。

S4090、特征提取子网络输出一个人脸特征向量。训练好后的特征提取子网络对同一个人的不同脸部图像会输出高度相似的特征向量，这个特征向量里包含了人的脸部相貌信息，可以用来识别出不同的人，也可以作为对长相的受欢迎程度进行预测的依据。

S4100、将 S4030 中解码得到的音频信息进行预处理之后，提取音频信息中的频谱信息。

S4110、将 S4030 中解码得到的音频信息进行预处理之后，提取音频信息中的梅尔倒谱系数。

视频中的背景音乐或者人物语音也是视频的重要内容之一，需要提取出背景音乐或者人物语音的音频信息的特征，根据音频信息的特征对视频的优质程度作一个预估。

S4120、将 S4100 和 S4110 中得到的音频信息和梅尔倒谱系数输入至音频特征提取网络。音频特征提取网络可以是任何基于 CNN 或者 RNN 的网络，预先在公开的用于音乐识别或人物语音识别的数据集上训练音频识别网络，再在优质视频训练集合上细调音频识别网络，得到音频特征提取网络。

S4130、音频特征提取网络输出一个音频特征向量。音频特征向量是指描述视频中声音信息的向量。

S4140、根据动作特征向量、人脸特征向量和音频特征向量构建视频特征向量。

S4150、将视频特征向量输入至视频评分网络。视频评估网络是可以是一个简单的多层感知器也可以是较为复杂的 RNN 网络。视频评估网络的输入是视频特征向量，输出是一个对视频优质程度的评分。

S4160、视频评分网络输出一个视频评分。所述视频评分用于衡量视频优质程度，评分越高，视频优质程度越高。

在上述实施例的基础上，本实施例提供一个实例。用户通过短视频类应用程序上传一个 1 分钟的短视频，一实施例中，短视频是人物 A 跳舞的视频，背景音乐为音乐 B。假设短视频中 1 秒中包含 24 帧。那么将 1 分钟的短视频进行解码，可以得到 1440 帧的图像数据和 1 分钟的音频信息。每帧图像数据中包含一张图像，即可以得到连续的 1440 张图像构成的图像序列。将由 1440 张图像构成的图像序列分别输入至动作特征提取网络和人脸特征提取网络，动作特征提取网络输出一个动作特征向量，该动作特征向量用于表征视频中人物 A 的舞蹈动作。人脸特征提取网络输出一个人脸特征向量，该人脸特征用于表征视频中人物 A 的脸部相貌信息。将音频信息输入至音频特征提取网络，音频特征提取网络输出一个音频特征向量。示例性的，动作特征向量是 3 行 5 列的向量，人脸特征向量是 5 行 8 列的向量，音频特征向量 4 行 9 列的向量，根据所述动作特征向量、人脸特征向量和音频特征向量构建视频特征向量；可以得到一个 5 行 22 列的视频特征向量，将视频特征向量输入至视频评估网络，视频评估网络输出一个评分，示例性的评分可以是 95 分。

本实施例提供的视频质量评估方法，首先通过对待评估视频进行解码，得到待评估视频的图像序列和音频信息，然后提取图像序列中的动作特征向量和人脸特征向量，以及提取音频信息中的音频特征向量，再根据动作特征向量、

人脸特征向量和音频特征向量构建视频特征向量；最后根据视频特征向量确定待评估视频的质量评分。可见，本实施例可以根据待评估视频中提取的动作特征向量、人脸特征向量和音频特征向量中的至少一个向量，确定待评估视频的质量评分，提高了视频质量评估的时效性，提升了优质视频评估的准确度。

实施例四

图 5 是本申请实施例四中的视频质量评估装置的结构示意图，本实施例可适用于评价视频质量的情况，该视频质量评估装置可以通过软件和/或硬件的方式实现。设定视频质量评估装置集成在终端设备中，该视频质量评估装置的结构如下：视频解码模块 510，设置为对待评估视频进行解码，得到所述待评估视频的图像序列和音频信息；向量提取模块 520，设置为提取所述图像序列中的动作特征向量和人脸特征向量，以及提取所述音频信息中的音频特征向量；向量构建模块 530，设置为根据所述动作特征向量、人脸特征向量和音频特征向量构建视频特征向量；质量评分模块 540，设置为根据所述视频特征向量确定所述待评估视频的质量评分。

本实施例提供的视频质量评估装置，首先通过对待评估视频进行解码，得到待评估视频的图像序列和音频信息，然后提取图像序列中的动作特征向量和人脸特征向量，以及提取音频信息中的音频特征向量，再根据动作特征向量、人脸特征向量和音频特征向量构建视频特征向量；最后根据视频特征向量确定待评估视频的质量评分。可见，本实施例可以根据待评估视频中提取的动作特征向量、人脸特征向量和音频特征向量中的至少一个向量，确定待评估视频的质量评分，提高了视频质量评估的时效性，提升了优质视频评估的准确度。

实施例五

图 6 为本申请实施例五提供的一种设备的结构示意图，如图 6 所示，该设备包括处理器 610 和存储器 620；设备中处理器 610 的数量可以是一个或多个，图 6 中以一个处理器 610 为例；设备的处理器 610 和存储器 620 可以通过总线或其他方式连接，图 6 中通过总线连接为例。

存储器 620 作为一种计算机可读存储介质，可设置为存储软件程序、计算机可执行程序以及模块，如本申请实施例中的视频质量评估方法对应的程序指令/模块（例如，视频质量评估装置中的视频解码模块 510、向量提取模块 520、向量构建模块 530 和质量评分模块 540）。处理器 610 通过运行存储在存储器 620 中的软件程序、指令以及模块，从而执行设备的多种功能应用以及数据处理，

即实现上述的视频质量评估方法。

实施例六

本申请实施例六还提供一种包含计算机可执行指令的存储介质，所述计算机可执行指令在由计算机处理器执行时用于执行一种视频质量评估方法，该方法包括：对待评估视频进行解码，得到所述待评估视频的图像序列和音频信息；提取所述图像序列中的动作特征向量和人脸特征向量，以及提取所述音频信息中的音频特征向量；根据所述动作特征向量、人脸特征向量和音频特征向量构建视频特征向量；根据所述视频特征向量确定所述待评估视频的质量评分。

本申请实施例所提供的一种包含计算机可执行指令的存储介质，该计算机可执行指令不限于如上所述的方法操作，还可以执行本申请任意实施例所提供的视频质量评估方法中的相关操作。

1、一种视频质量评估方法，其中，包括：

对待评估视频进行解码，得到所述待评估视频的图像序列和音频信息；

提取所述图像序列中的动作特征向量和人脸特征向量，以及提取所述音频信息中的音频特征向量；

根据所述动作特征向量、所述人脸特征向量和所述音频特征向量构建视频特征向量；

根据所述视频特征向量确定所述待评估视频的质量评分。

2、根据权利要求 1 所述的方法，其中，所述提取所述图像序列中的动作特征向量包括：

利用动作特征提取网络对所述图像序列中每一帧图像进行特征提取，得到所述图像序列中的动作特征向量；

或，

提取所述图像序列中的每一帧图像的光流数据；

利用动作特征提取网络对所述图像序列中每一帧图像的光流数据进行特征提取，得到所述图像序列中的动作特征向量。

3、根据权利要求 2 所述的方法，还包括：

获取训练视频集合，并提取所述训练视频集合中每个训练视频中的图像序列；

在视频数据库中训练出动作分类网络；

使用所述训练视频集合中每个训练视频中的图像序列对所述动作分类网络进行训练，并删除所述动作分类网络中的分类层，得到所述动作特征提取网络。

4、根据权利要求 1 所述的方法，其中，所述提取所述图像序列中的人脸特征向量包括：

利用人脸特征提取网络提取所述图像序列中的人脸特征向量。

5、根据权利要求 4 所述的方法，其中，所述人脸特征提取网络包括人脸检测子网络和特征提取子网络。

6、根据权利要求 5 所述的方法，其中，所述利用人脸特征提取网络提取所述图像序列中的人脸特征向量包括：

利用所述人脸检测子网络检测出所述图像序列中每一帧图像中的人脸图像；

利用所述特征提取子网络对所述图像序列中每一帧图像中的人脸图像进行特征提取，得到所述图像序列中的人脸特征向量。

7、根据权利要求 5 或 6 所述的方法，还包括：

获取训练视频集合，并提取所述训练视频集合中每个训练视频中的人脸图像；

在人脸数据库中训练出人脸识别网络；

使用所述训练视频集合中每个训练视频中的人脸图像对所述人脸识别网络进行训练，得到所述特征提取子网络。

8、根据权利要求 1 所述的方法，其中，所述提取所述音频信息中的音频特征向量包括：

提取所述音频信息的倒谱系数和频谱中的至少之一；

利用音频特征提取网络对所述倒谱系数和所述频谱中的至少之一进行特征提取，得到音频特征向量。

9、根据权利要求 8 所述的方法，还包括：

获取训练视频集合，并提取所述训练视频集合中每个训练视频中的音频信息；

在音频数据库中训练出音频识别网络；

使用所述训练视频集合中每个训练视频中的音频信息对所述音频识别网络进行训练，得到所述音频特征提取网络。

10、根据权利要求 1-9 中任一项所述的方法，其中，所述根据所述视频特征向量确定所述待评估视频的质量评分包括：

将所述视频特征向量输入到视频评估网络中进行评估，得到所述待评估视频的质量评分。

11、根据权利要求 10 所述的方法，其中，所述视频评估网络包括第一层评估网络、第二层评估网络和输出层。

12、根据权利要求 11 所述的方法，其中，所述将所述视频特征向量输入到视频评估网络中进行评估，得到所述待评估视频的质量评分包括：

将所述视频特征向量输入所述第一层评估网络，第一层评估网络通过递归和非线性运算获得初级特征向量；

所述第一层评估网络将所述初级特征向量输入第二层评估网络和输出层，所述第二层评估网络通过递归和非线性运算获得高级特征向量并将所述高级特

征向量输入所述输出层;

所述输出层对所述初级特征向量和所述高级特征向量进行运算,得到所述待评估视频的质量评分。

13、一种视频质量评估装置,包括:

视频解码模块,设置为对待评估视频进行解码,得到所述待评估视频的图像序列和音频信息;

向量提取模块,设置为提取所述图像序列中的动作特征向量和人脸特征向量,以及提取所述音频信息中的音频特征向量;

向量构建模块,设置为根据所述动作特征向量、所述人脸特征向量和所述音频特征向量构建视频特征向量;

质量评分模块,设置为根据所述视频特征向量确定所述待评估视频的质量评分。

14、一种设备,包括:

至少一个处理器;

存储器,设置为存储至少一个程序;

当所述至少一个程序被所述至少一个处理器执行,使得所述至少一个处理器实现如权利要求 1-12 中任一所述的视频质量评估方法。

15、一种计算机可读存储介质,存储有计算机程序,所述程序被处理装置执行时实现如权利要求 1-12 中任一所述的视频质量评估方法。