(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104142995 A
(43)申请公布日 2014.11.12

(21)申请号 201410370304.7

(22)申请日 2014.07.30

(71)申请人 中国科学院自动化研究所
地址 100190 北京市海淀区中关村东路 95 号

(72)发明人 徐常胜 杨小汕 张天柱

(74)专利代理机构 中科专利商标代理有限责任公司 11021
代理人 宋焰琴

(51)Int.Cl.
G06F 17/30(2006.01)

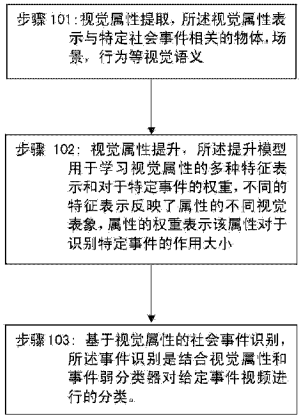
权利要求书3页 说明书7页 附图1页

(54)发明名称

基于视觉属性的社会事件识别方法

(57)摘要

本发明公开了一种基于视觉属性的社会事件识别方法，该方法包括：步骤101，从训练事件视频集中的视频中提取视觉属性，所述视觉属性用于描述与视频对应的事件相关的属性；其中，所述训练事件视频中的每个视频对应一个已知事件类别；步骤102，基于提升的迭代模型，训练得到视觉属性分类器，并利用所述视觉属性分类器训练得到多个事件弱分类器；步骤103，基于上述得到的视觉属性分类器以及多个事件弱分类器对待分类事件视频进行分类。本发明针对传统的基于属性的视频事件识别方法中需要大量人工给定的语义标签问题，提出了自动的视觉属性挖掘方法；另外针对视频事件识别中视觉属性复杂多变的问题，本发明对同一种视觉属性建立了多种特征表示。



CN 104142995 A

1. 一种基于视觉属性的社会事件识别方法，其特征在于，该方法包括以下步骤：

步骤 101，从训练事件视频集中的视频中提取视觉属性，所述视觉属性用于描述与视频对应的事件相关的属性；其中，所述训练事件视频中的每个视频对应一个已知事件类别；

步骤 102，基于提升的迭代模型，训练得到视觉属性分类器，并利用所述视觉属性分类器训练得到多个事件弱分类器；

步骤 103，基于上述得到的视觉属性分类器以及多个事件弱分类器对待分类事件视频进行分类。

2. 根据权利要求 1 所述的方法，其特征在于，所述步骤 101 进一步包括以下步骤：

步骤 1011，从所述训练事件视频集中每个视频的文本描述中提取语义单词和词组；

步骤 1012，收集所述语义单词和词组对应的图像，根据视觉信息计算语义单词和词组的视觉表示力，结合语义单词或词组的语义粘滞性，从语义单词和词组中选出多个视觉属性。

3. 根据权利要求 2 所述的方法，其特征在于，设De为一个视频的文本描述，De被分割为多个语义或者词组单元$De=<\mathbf{se}_1,\mathbf{se}_2,...,\mathbf{se}_m>$，其中 se_i 表示一个语义单元；视频的文本描述分割问题可以进一步表示为一个优化问题：$\arg\max_{\mathbf{se}_1,...,\mathbf{se}_m} Stc(De)$；

这里$Stc(De)=\sum_{i=1}^{m} Stc(\mathbf{se}_i)$，其中 Stc 表示衡量分割词组粘滞性的函数。

4. 根据权利要求 2 所述的方法，其特征在于，一个分割词组 se 被选为视觉属性的概率是由 se 的语义粘滞性和视觉表示力共同决定的：

$Score(se) = Stc(se)V_{flickr}(se)$

这里 V_{flickr} 是 se 的视觉表示力，是通过收集得到的与其对应的图像集的视觉相似性来计算得到：

$$V_{flickr}(\mathbf{se})=\sum_{i\in I_{se}} sim(i,Cent(I_{se}))$$

其中，I_{se} 是当 se 作为检索词时，从图像共享网站搜索得到的图像集；$Cent(I_{se})$ 表示 I_{se} 的重心；sim() 表示图像的相似度。

5. 根据权利要求 1 所述的方法，其特征在于，步骤 102 具体通过迭代执行以下三个步骤：

步骤 1021：学习领域适应的共有特征表示，该步骤中利用权重分布采样所有视频对应的图像帧集合和辅助图像集中的图像；并利用去噪自编码器学习所采样的样本图像的共有特征表示；其中，所述辅助图像集是利用所述训练事件视频集中所有视频对应的已知事件类别名称作为关键词从图像共享网站检索得到；

步骤 1022：利用所学习得到的领域适应的共有特征表示，训练得到与所述视觉属性对应的多个属性分类器，并利用所述属性分类器更新所述视频集中所有视频对应的图像帧以及辅助图像集中图像的权重；

步骤 1023：利用所有视觉属性分类器对训练视频集中的每个视频进行描述以构造每个视频的视觉属性特征向量，然后利用视觉属性特征向量训练得到事件视频弱分类器；，

训练得到事件视频弱分类器，并利用所述事件视频弱分类器进一步更新训练视频集中所有视频对应的图像帧的权重。

6. 如权利要求5所述的方法，其特征在于，步骤1021中，去噪自编码器利用加了噪声后的特征恢复得到原来的特征，其重构误差如下表示：

$$\mathcal{L}_{sq}(\mathbf{W})=\frac{1}{2sr}\sum_{j=1}^{r}\sum_{i=1}^{s}\|\mathbf{x}_i-\mathbf{W}\tilde{\mathbf{x}}_{ij}\|^2$$

其中，$\mathcal{L}_{sq}(\mathbf{W})$是指去噪自编码器的重构误差，w表示将所述训练视频集中所有视频对应的图像帧和辅助图像集中的图像帧映射成共有特征表示的映射矩阵；s表示采样得到的样本个数，r表示对每个样本加噪声的次数；x_i是第i个样本的原始特征，$\tilde{\mathbf{x}}_{ij}$是对第i个样本的原始特征第j次加噪声以后的特征；

通过上述重构误差方程可以求得映射矩阵W的解析解，具体如下表示：

$$\mathbf{W}=E[\mathbf{P}]E[\mathbf{Q}]^{-1},\mathbf{Q}=\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\top},\mathbf{P}=\bar{\mathbf{X}}\tilde{\mathbf{X}}^{\top} \quad (9)$$

其中，$X=[x_1,\ldots,x_s]$表示采样得到的样本集合，$\bar{\mathbf{X}}=[\mathbf{X},\ldots,\mathbf{X}]$，另外$\tilde{\mathbf{X}}$是由$\bar{\mathbf{X}}$加噪声后的特征向量组成；E表示期望。

7. 如权利要求6所述的方法，其特征在于，步骤1022中，属性分类器的分类误差如下表示：

$$\epsilon^c=\frac{1}{\sum_{i\in image(c)}d_i}\sum_{i\in image(c)}d_i\cdot\mathbb{I}\left(1\neq\mathbf{f}^c(g(\mathbf{x}_i))\right) \quad (10)$$

其中，ϵ^c表示分类误差，$\mathbb{I}$表示符号函数，如果括弧中的条件满足，则函数值为1，否则函数值为0；x_i表示第i个样本的特征，即$\mathbf{X}_v=\{\mathbf{x}_i\}_{i=1}^{I}$中的第i个样本的特征向量；$g(x_i)$表示将$x_i$的特征映射为共有特征表示后的特征；$f^c(g(x_i))$表示第c个属性分类器；

利用训练得到的分类器如下更新视频集和辅助图像集中图像的权重：

$$d_i=d_i\exp\left(\alpha^c\mathbb{I}\left(1\neq\mathbf{f}^c(g(\mathbf{x}_i))\right)\right),\forall i\in image(c)$$

$$\alpha^c=\ln\left((1-\epsilon^c)/\epsilon^c\right)$$

其中，d_i表示第i个图像的权重，image(c)表示第c个属性分类器对应的视频包含的所有帧图像；α^c表示权重更新率。

8. 如权利要求7所述的方法，其特征在于，步骤1023中每个视频的所述视觉属性特征向量如下构建：

利用所有属性分类器对所述训练视频集中每个视频对应的图像帧得到分类输出值，这些分类输出值构成图像帧的视觉属性特征向量，将一个视频对应的所有帧图像对应的视觉属性特征向量进行池化得到该视频的视觉属性特征向量。

9. 如权利要求5所述的方法，其特征在于，步骤1023中，事件视频弱分类器的分类误差和权重如下计算：

$$\epsilon = \frac{1}{\sum_{j=1}^{n} \hat{d}_j} \sum_{j=1}^{n} \hat{d}_j \cdot \mathbb{I}\left(y_j \neq \mathbf{h}(\mathbf{v}_j)\right)$$

$$\alpha = \ln\left((1-\epsilon)/\epsilon\right) + \ln(K-1)$$

其中，ϵ表示事件视频弱分类器的分类误差，v_j表示第 j 个视频，y_j表示训练事件视频集中第 j 个视频的事件类别；$h(v_j)$表示对视频 v_j 训练得到的事件视频弱分类器，α 表示事件视频弱分类器 $h(v_j)$ 的权重；$\hat{d}_j$表示第 j 个视频的权重，K 表示事件类别的个数。

10. 如权利要求 8 所述的方法，其特征在于，步骤 103 具体包括：

对于待识别视频，利用映射矩阵 W 计算其对应的图像帧的特征表示；

将所述特征表示作为所述属性分类器的输入，进而得到待识别视频的视觉属性特征向量；

将所述待识别视频的视觉属性特征向量作为所有事件视频弱分类器的输入，对所述待识别视频进行分类。

基于视觉属性的社会事件识别方法

技术领域

[0001] 本发明属于社会媒体(social media)挖掘和视频分析领域，具体涉及基于图像分享网站和视频分享网站的视觉属性的社会事件的识别方法。

背景技术

[0002] 随着手机、数字摄像头以及Flickr、Youtube等社交媒体的不断普及，人们变得更容易从网络上获取和分享信息。这使得发生在人们周围的社会事件以更快的速度传播并随之产生了大量与事件相关的不同模态的媒体数据，例如图像、文本和视频。根据大量多媒体数据来理解特定社会事件可以更好地帮助人们浏览、搜索和监控社会事件。但由于社会事件的复杂多变，如何有效地挖掘媒体数据来理解社会事件仍然是一个难题。

[0003] 近年来，已有大量利用各种媒体数据的社会事件识别和检测的方法被提出。针对MediaEval公布的多媒体事件检测问题，图像的文本描述，标签、地理位置以及时间标记等数据被广泛用于事件的理解与检测。这些方法所关注的社会事件是发生在特定时间、地点的一类事件，例如“发生在西班牙巴塞罗拉和意大利罗马的所有足球事件”。还有一些方法借助社交网站、博客、维基以及搜索引擎中的大量文本信息来挖掘更为抽象的社会事件，例如“拉里·佩奇和谢尔盖·布林在1998年创立了谷歌公司”。除此之外，还有大量的方法被提出用于检测和识别视频中的事件。例如在多媒体事件检测(MED)数据集中，视频事件主要是关于“生日聚会”，“做蛋糕”以及“攀岩”等。由于包含在图像和视频中的视觉语义信息不易被提取和利用，目前的事件识别方法难以在视频事件中获得好的效果。为了改进对视频的社会事件的理解和识别，目前有大量的方法依赖于属性来描述视频中的事件。

[0004] 目前基于属性的视频事件识别方法可以分为三个主要步骤。(1)人工标定视觉样本(图像或视频)的属性，这些属性是人为选定的最能体现事件特征的语义信息。(2)利用包含属性标记的视频或图像样本训练属性分类器。(3)利用属性分类器进一步得到视频的属性描述特征向量。最终将根据视频的属性描述特征向量来进行事件分类。尽管目前基于属性的方法可以得到好的效果，但仍然存在大量问题。一方面是标定属性需要耗费大量人力成本。另一方面是给定属性对应的单个分类器不足以描述事件对应的复杂多变的视觉外观。

发明内容

[0005] 本发明的目的是通过自动挖掘视觉属性，得到对视频中的事件更有效的特征描述方式，进而可以得到更好的分类效果。针对事件复杂多变的视觉外观，用多种特征来描述给定的视觉属性，可以更全面的表达事件的视觉外观。

[0006] 为实现上述目的，本发明提供一种基于视觉属性的社会事件识别方法，该方法包括以下步骤：

[0007] 步骤101，从训练事件视频集中的视频中提取视觉属性，所述视觉属性用于描述与视频对应的事件相关的属性；其中，所述训练事件视频中的每个视频对应一个已知事件类

别；

[0008] 步骤 102，基于提升的迭代模型，训练得到视觉属性分类器，并利用所述视觉属性分类器训练得到多个事件弱分类器；

[0009] 步骤 103，基于上述得到的视觉属性分类器以及多个事件弱分类器对待分类事件视频进行分类。

[0010] 本发明的有益效果：本发明通过自动挖掘视觉属性，减少了传统基于视觉属性的事件识别方法中需要人工标定属性的耗费。基于提升的多特征属性表示方法可以有效地表示视频事件中复杂多变的视觉外观。

附图说明

[0011] 图 1 是本发明基于视觉属性的社会事件识别方法的流程图；

具体实施方式

[0012] 为使本发明的目的、技术方案和优点更加清楚明白，以下结合具体实施例，并参照附图，对本发明进一步详细说明。

[0013] 图 1 为本发明提出的基于视觉属性的社会事件识别方法的流程图，所述方法通过自动挖掘视觉属性得到可以识别社会事件的关键视觉属性，这些视觉属性被进一步提升来更好地表示社会事件，最终视频事件被表示为视觉属性的特征向量。如图 1 所示，所述方法包括三个部分：1) 视觉属性提取，2) 视觉属性提升，3) 基于视觉属性的社会事件识别。具体来说，所述方法包括以下步骤：

[0014] 步骤 101，视觉语义属性提取，所述视觉语义属性表示描述特定事件相关的物体，场景，行为等视觉语义；物体可以是人、车或者动物等；场景可能是体育场、教堂等，行为主要是人的行为活动，比如拥抱、握手等。

[0015] 所述步骤 101 进一步包括以下步骤：

[0016] 步骤 1011，从训练事件视频集中的每一个事件视频的文本描述中提取语义单词和词组；其中，所述训练事件视频集中的每一个事件视频对应一个特定的社会事件，即每个事件视频具有一个事件类别；所述训练事件视频集中的所有事件视频对应预定数目个社会事件，所述预定数目小于训练视频集中的视频个数。

[0017] 设$\mathcal{De}$为特定社会事件相关的一个事件视频的文本描述，$\mathcal{De}$可以被分割为多个语义或者词组单元$\mathcal{De}=<\mathbf{se}_1,\mathbf{se}_2,...,\mathbf{se}_m>$，其中 se_i 表示一个语义或词组单元，所述词组单元为最能表达视觉属性的词组。视频的文本描述分割问题可以进一步表示为一个优化问题：

[0018]

$$\arg\max_{\mathbf{se}_1,...,\mathbf{se}_m} Stc(\mathcal{De}) \quad (1)$$

[0019] 这里$Stc(\mathcal{De})=\sum_{i=1}^{m} Stc(\mathbf{se}_i)$，其中 Stc 表示衡量分割词组粘滞性的函数。较高的粘滞值表示词组被进一步分割会影响词组的语义完整性。明确来说，Stc 被定义为

[0020] $Stc(se) = L(se)e^{Q(se)}Sigmod(SCP(se))$ (2)

[0021] 这里 Q(se) 表示 se 作为关键词条的概率，如出现在维基百科中的概率等。SCP 表示 N 元文法模型的对称条件概率，根据 se 所有可能的二分方式，SCP 可以被定义为：

[0022] $$SCP(\mathbf{se}) = log \frac{Pr(\mathbf{se})^2}{\frac{1}{n-1}\sum_{i=1}^{n-1} Pr(w_1,\ldots,w_i)Pr(w_{i+1}\ldots w_n)} \quad (3)$$

[0023] 这里 Pr() 表示一先验概率，即为括号内词组序列的联合概率，该联合概率根据该词组序列中每个词组的条件概率乘积得到。n 表示 se 中单词的个数，w 表示 se 中的某个单词。所述 Pr() 可以直接由微软 N 元文法服务得到，所述微软 N 元文法服务是一个开源的云计算项目，用户可以给该服务的服务器发送一个词组，该云服务就可以返回该词组序列的联合概率。L(se) 被用来优先选择较短的词组分割结果，se 的绝对值表示词组 se 中的单词个数。

[0024] $$L(\mathbf{se}) = \begin{cases} \frac{(|\mathbf{se}|-1)}{|\mathbf{se}|}, for\ |\mathbf{se}|>1 \\ 1, for\ |\mathbf{se}|=1. \end{cases} \quad (4)$$

[0025] 其中，|se| 表示 se 中的单词个数。

[0026] 步骤 1012，收集语义单词和词组 se 对应的图片，这里的图片可以是用步骤 1011 中得到的词组 se 作为检索词时，从 Flickr 上返回的图片；根据视觉信息计算语义单词和词组 se 的视觉表示力，结合语义单词或词组的语义粘滞性，从语义单词和词组中选出视觉语义属性集，即选出视觉语义属性概率较高的预定数量的语义单词或词组。实验证明使用 500 个左右的词组就可以达到最好的事件识别效果。

[0027] 一个分割词组 se 被选为视觉语义属性的概率是由 se 的语义粘滞性和视觉表示力共同决定的。

[0028] $Score(se) = Stc(se)V_{flickr}(se)$ (5)

[0029] 这里 V_{flickr} 是 se 的视觉表示力，是通过图像集的视觉相似性来计算得到：

[0030] $$V_{flickr}(\mathbf{se}) = \sum_{i \in I_{se}} sim(i, Cent(I_{se})) \quad (6)$$

[0031] 这里的 I_{se} 是当 se 作为检索词时，图像共享网站 Flickr 上返回的图像集。可选地为每个 se 从 Flickr 上收集了大约 100 张图像。$Cent(I_{se})$ 表示 I_{se} 的重心。图像的重心是指图像对应的特征向量的重心。图像集 I_{se} 的重心通过对图像集 I_{se} 中所有图像的特征向量求均值来计算得到。假设 v_i 和 v_j 是图像 i 和 j 的特征向量，这里的图像相似度 sim() 是借助于傅里叶变换来计算得到，这里 $\mathcal{F}$ 表示傅里叶变换，λ 是一个正则化参数，预先设定的，用于防止分母太小的时候而计算得到无意义的相似性值。

[0032]

$$sim(i,j) = \mathcal{F}^{-1}\left(\frac{\mathcal{F}(\mathbf{v}_i)^* \odot \mathcal{F}(\mathbf{v}_j)}{\mathcal{F}(\mathbf{v}_i)^* \odot \mathcal{F}(\mathbf{v}_i) + \lambda}\right) \quad (7)$$

[0033] 步骤 102，视觉语义属性提升，即通过提升模型对视觉语义属性进行提升。所述提升模型用于学习视觉语义属性集中视觉语义属性的多种特征表示和对于特定事件的权重，

不同的特征表示反映了属性的不同视觉表象，属性的权重表示该属性对于识别特定事件的作用大小；即该步骤中基于提升 (boosting) 的迭代模型

[0034] 所述步骤 102 是基于提升 (boosting) 的迭代模型来构建的：

[0035] 符号假设：$\mathcal{V}=\{\mathbf{v}_j\}_{j=1}^{n}$ 表示所述训练事件视频集，其包括 n 个事件视频。这里 v_j 表示一个由 l_j 帧图像组成的事件视频。$\mathbf{Y}_{\mathcal{V}}=\{y_j\}_{j=1}^{n}, y_j\in\{1,2,\ldots,K\}$ 表示 $\mathcal{V}$ 中所有视频对应的事件类别，即不同的社会事件，K 是视频所包含的事件类别的个数。$l=\sum_{j=1}^{n} l_j$ 表示视频中所有帧图像的总数。$\mathbf{X}_{\mathcal{V}}=\{\mathbf{x}_i\}_{i=1}^{l}$ 表示视频集中所有帧图像的视觉特征向量。我们用 Cpts 表示步骤 101 中提取得到的视觉语义属性的集合。$\mathbf{A}_{\mathcal{V}}=\{\mathbf{a}_i\}_{i=1}^{l}$ 表示视频集中所有帧图像的视觉语义标签，一个帧图像的视觉语义标签为该帧图像所属的事件视频的视觉语义属性集合。这些视觉语义标签可以根据 101 步骤中得到的视频的视觉语义属性得到。在步骤 101 中提取视觉语义属性的过程中，每个视觉语义属性都是从某个视频的文本描述中提取，因此可以给视频自动赋予视觉语义属性标签。对于某个视频 v_j，如果 $c\in Cpts$ 是从它的文本信息中提取出来的视觉语义属性，那么视频 v_j 中的所有图像的视觉语义属性标签都会包含 c。另外假设 $\mathcal{I}$ 为一个辅助图像集，我们用 $\mathbf{X}_{\mathcal{I}}=\{\mathbf{x}_i\}_{i=l+1}^{l+m}$ 表示所述辅助图像集中所有图像的视觉特征向量。这个辅助图像集是以每个事件类别的名称作为检索词从 Flickr 上得到。所述辅助图像集中所有图像的视觉语义标签表示为 $\mathbf{A}_{\mathcal{I}}=\{\mathbf{a}_i\}_{i=1}^{m}$，辅助图像集中的图像都是从 Flickr 上检索得到，因此都带有一定的文本描述信息。如果一幅图像 x_i 的文本中包含有某个视觉语义属性 $c\in Cpts$，那么这个图像的视觉语义属性标签 a_i 就包含这个视觉语义属性 c。另外我们假设训练事件视频集和辅助图像集中的所有图像的权重分布表示为 $\mathbf{d}^{V}=\{d_i\}_{i=1}^{l}$ 和 $\mathbf{d}^{I}=\{d_i\}_{i=l+1}^{l+m}$，这些权重将在提升迭代中不断被更新。

[0036] 在每次提升迭代中，我们首先根据 d^V 和 d^I 来学习领域适应的特征表示，然后利用这些特征来训练视觉属性分类器，视觉属性分类器将进一步被用于训练社会事件的事件识别分类器。在所述视觉属性的提升迭代过程中涉及的领域适应的特征学习，视觉属性分类器训练，利用视觉属性的社会事件识别三个主要部分将分别在下面的步骤 1021，步骤 1022 和步骤 1023 中介绍。

[0037] 步骤 1021 领域适应的特征学习；

[0038] 为了同时利用视频集中的帧图像和辅助图像集中的图像，我们需要学习没有领域差异的特征表示方式。这里我们采用边缘化的去噪自编码器 (mSDA) 来学习视频帧图像和辅助图像集中的图像的共有的特征表示。假设 $\{\mathbf{x}_i\}_{i=1}^{s}$ 是从视频帧图像和辅助图像集中的图像采样得到的样本图像的原始特征向量，采样是根据权重分布 d^V 和 d^I 从所有图像样本中选取预定数目如 1/10 的图像样本。$\tilde{\mathbf{x}}_i$ 是 x_i 的加噪声后的特征向量，即随机将 x_i 中的某些元素置为 0，利用去噪自编码器可以利用没有被噪声干扰的数据恢复出丢失的数据。mSDA 方法用单个映射函数构造平方优化目标方程来重建原始特征向量。在 mSDA 中，把视频帧图像和辅助图像集中的图像放在一起训练可以减小领域差异。通常对特征向量加多重噪声，此时优化目标方程为：

[0039]

$$\mathcal{L}_{sq}(\mathbf{W})=\frac{1}{2sr}\sum_{j=1}^{r}\sum_{i=1}^{s}\|\mathbf{x}_i-\mathbf{W}\tilde{\mathbf{x}}_{ij}\|^2 \quad (8)$$

[0040] 这里我们采用的去噪自编码器是用加了噪声以后的特征恢复得到原来的特征。$\mathcal{L}_{sq}(\mathbf{W})$是指去噪自编码器的重构误差，也就是恢复得到的特征与原始特征的误差。w 表示映射矩阵，x_i 是第 i 个样本图像的原始特征，$\tilde{\mathbf{x}}_{ij}$是对第 i 个样本的原始特征第 j 次加噪声以后的特征。s 表示训练样本个数，r 表示对每个样本加噪声的次数。

[0041] 这个二次优化方程可以求得解析解：

[0042]

$$\mathbf{W}=E[\mathbf{P}]E[\mathbf{Q}]^{-1},\mathbf{Q}=\tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\top},\mathbf{P}=\bar{\mathbf{X}}\tilde{\mathbf{X}}^{\top} \quad (9)$$

[0043] 这里 $X=[x_1,\ldots,x_s]$，$\bar{\mathbf{X}}=[\mathbf{X},\ldots,\mathbf{X}]$，另外$\tilde{\mathbf{X}}$是由$\bar{\mathbf{X}}$加噪声后的特征向量组成。此外 E 表示期望，T 表示矩阵的转置，-1 表示矩阵的逆。

[0044] 所述共有特征表示就是将原有特征乘以所述映射矩阵 W 得到。在下面的步骤中，用 $g(x_i)$ 表示原来的特征向量 x_i 被所述映射矩阵 W 映射后的特征。

[0045] 步骤 1022 训练视觉属性分类器

[0046] 基于上一步得到的共有特征表示，我们学习属性分类器。每个属性分类器都是用线性 SVM 训练得到。考虑到只有视频的某些帧图像才与属性相关联，因此我们只是采样了视频中的一部分帧图像来训练分类器。这里是用步骤 1021 中的图像样本子集的共有特征表示$\{g(\mathbf{x}_i)\}_{i=1}^{s}$来训练属性分类器。训练得到属性分类器之后，我们可以根据分类准确率更新视频帧图像的权重。属性分类器准确分类的那些帧图像样本将在下次 boosting 迭代中更容易被选择出来用于训练属性分类器。对于第 c 个属性分类器，分类误差与权重更新率可以分别计算为：

[0047]

$$\epsilon^c=\frac{1}{\sum_{i\in image(c)}d_i}\sum_{i\in image(c)}d_i\cdot\mathbb{I}\left(1\neq\mathbf{f}^c(g(\mathbf{x}_i))\right) \quad (10)$$

[0048] ϵ^c表示分类误差，$\mathbb{I}$表示符号函数，如果括弧中的条件满足，则函数值为 1，否则函数值为 0；x_i 表示第 i 个训练样本的特征，即$\mathbf{X}_V=\{\mathbf{x}_i\}_{i=1}^{I}$中的第 i 个图像帧的特征向量；$g(x_i)$ 表示 x_i 对应的步骤 1021 中所述的共有特征表示；$f^c(g(x_i))$ 表示第 c 个属性分类器，如果 x_i 是属于属性 c，则函数值为 1，否则为 0，所述属性 c 就是前面提取得到的视觉语义属性；d_i 是第 i 个训练样本权重，image(c) 表示属性 c 对应的所有图像，包括提取出视觉语义属性 c 的视频包括的所有帧图像。

[0049]

$$\alpha^c=\ln\left((1-\epsilon^c)/\epsilon^c\right) \quad (11)$$

[0050] α^c 表示权重更新率。对视频中的所有帧图像，其权重更新方式可以表示为：

[0051]

$$d_i = d_i \exp\left(\alpha^c \mathbb{I}\left(1 \neq \mathbf{f}^c(g(\mathbf{x}_i))\right)\right), \forall i \in image(c) \quad (12)$$

[0052] 其中 image(c) 表示第 c 个属性分类器对应的视频包含的所有帧图像。试验中我们为了得到最好的效果选择 500 个左右视觉语义属性。每次迭代中每个视觉语义属性都对应一个属性分类器。在步骤 101 提取视觉语义属性的过程中，每个视觉语义属性都是从某个视频的文本描述中提取，因此可以给视频自动赋予视觉语义属性标签。对于帧图像来说，帧所在的视频属于哪个视觉语义属性，这个帧图像也就具有和视频相同的视觉语义属性。用同样的方式，更新辅助图像集中所有图像的权重。

[0053] 步骤 1023 利用视觉属性分类器训练得到事件视频的分类器，并更新视频帧图像的权重；

[0054] 根据步骤 1022，我们可以利用视觉属性分类器来描述事件视频。这里说的描述事件视频主要是指对每个视频得到一个特征表示向量。在步骤 1022 中，我们最终是得到了每个视觉语义属性对应的属性分类器。用所有属性分类器对视频进行打分，也就是用所有属性分类器的输出构造一个视频的特征描述。具体来说，事件视频中的每个帧图像都可以利用视觉属性分类器得到一个分类输出值。这些分类器的输出构成一个关于视觉属性的特征向量。然后我们把视频中所有帧图像对应的视觉属性特征向量做池化得到视频的视觉属性特征向量。池化简单来说就是把多个特征向量变成一个特征向量，但需要尽量保持原来特征向量所表达的信息。实现的时候我们是对多个特征向量的每一维取最大值，得到一个特征向量。最后我们利用这些视觉属性特征向量进行事件识别，进而训练得到事件视频的分类器。这里我们根据事件的识别准确率对视频对应的帧图像权重做进一步调整。根据视频帧图像样本的权重，我们可以计算关于视频的权重向量。

[0055] $$\hat{d}_j = \sum_{i \in image(j)} d_i, \forall j = 1, \ldots, n \quad (13)$$

[0056] 其中$\hat{d}_j$表示第 j 个视频的权重，d_i 是图像 i 的权重，image(j) 表示第 j 个视频对应的所有帧图像。n 表示视频总数。

[0057] 社会事件视频分类误差与事件视频弱分类器权重可以分别计算为：

[0058]

$$\epsilon = \frac{1}{\sum_{j=1}^{n} \hat{d}_j} \sum_{j=1}^{n} \hat{d}_j \cdot \mathbb{I}\left(y_j \neq \mathbf{h}(\mathbf{v}_j)\right) \quad (14)$$

[0059]

$$\alpha = \ln\left((1-\epsilon)/\epsilon\right) + \ln(K-1) \quad (15)$$

[0060] 这里 v_j 表示第 j 个视频，前面符号假设部分有说明，y_j 表示训练事件视频集中第 j 个视频的事件类别。$h(v_j)$ 表示对视频 v_j 训练得到的事件弱分类器，α 表示弱分类器 $h(v_j)$ 的权重。这里的事件弱分类器是根据步骤 1023 池化得到的视频的视觉属性特征向量来训练得到的。$\mathbb{I}$表示符号函数，如果括弧中的条件满足，则函数值为 1，否则函数值为 0。K 表示事件类别的个数。

[0061] 所有视频帧图像的权重可以相应被更新，l 表示所有视频的所有帧图像总数：

[0062]

$$d_i = d_i \cdot \exp\left(\alpha \cdot \mathbb{I}\left(y_j \neq \mathbf{h}(\mathbf{v}_j)\right)\right), i \in image(j), \forall i = 1,\ldots,l \quad (16)$$

[0063] 上述三个步骤 1021，1022 和 1023 不断迭代进行，帧图像和辅助数据集的权重不断被更新，每次迭代都会产生一个新的特征表示方式，一个事件分类器，每个属性都产生一个属性分类器。因此经过 T 次迭代后，对于得到 T 个特征表示，T 个事件分类器，对每个属性也得到 T 个属性分类器。

[0064] 步骤 103，基于视觉属性的社会事件识别，所述社会事件识别是根据视觉属性的图像样本权重和视觉属性分类器来识别特定社会事件。在步骤 102 中，随着提升过程的不断迭代，我们得到了视觉属性的多种特征表示和多个属性分类器。同时我们也得到了特定事件相关的视觉属性的权重，以及每个视觉属性对应的帧图像的权重。得到这些特征表示和事件分类器之后，我们就可以构造用于识别特定社会事件相关的视频 v 的分类器 H(v)，这里 α_i 表示第 t 次迭代中产生的弱分类器 $h_t(v)$ 的权重，由于总共有 K 个事件类别，因此 k 的取值范围是 1 到 K。$\mathbb{I}\left(\mathbf{h}_t(\mathbf{v}) = k\right)$是符号函数，如果弱分类器 $h_t(v)$ 的输出为 k，那么$\mathbb{I}(\cdot)$函数的输出为 1，否则为 0。

[0065]

$$\mathbf{H}(\mathbf{v}) = \arg\max_k \sum_{t=1}^{T} \alpha_t \cdot \mathbb{I}\left(\mathbf{h}_t(\mathbf{v}) = k\right), k \in \{1,\ldots,K\} \quad (17)$$

[0066] 具体来说，对于某个测试视频 v，第 t 个弱分类器 $h_t(v)$ 是按如下方式进行分类的：我们先利用步骤 1021 中学习得到的映射矩阵 w 来计算新的特征表示，然后采用步骤 1022 中的属性分类器得到视觉属性特征向量，最后采用步骤 1023 中的事件弱分类器 h(v) 对其分类。最终测试视频 v 的事件类别是根据 (17) 式所示的方式由 T 个弱分类器的结果共同决定。

[0067] 以上所述的具体实施例，对本发明的目的、技术方案和有益效果进行了进一步详细说明，所应理解的是，以上所述仅为本发明的具体实施例而已，并不用于限制本发明，凡在本发明的精神和原则之内，所做的任何修改、等同替换、改进等，均应包含在本发明的保护范围之内。

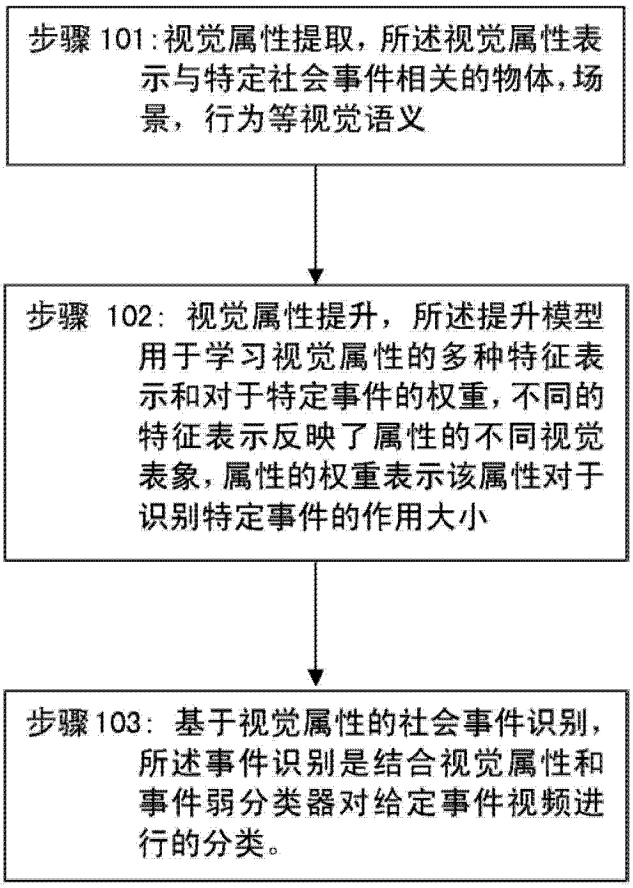

步骤 101:视觉属性提取，所述视觉属性表示与特定社会事件相关的物体，场景，行为等视觉语义
步骤 102：视觉属性提升，所述提升模型用于学习视觉属性的多种特征表示和对于特定事件的权重，不同的特征表示反映了属性的不同视觉表象，属性的权重表示该属性对于识别特定事件的作用大小
步骤 103：基于视觉属性的社会事件识别，所述事件识别是结合视觉属性和事件弱分类器对给定事件视频进行的分类。

图 1