基于相关核映射线性近邻传播的视频语义标注*

张建明，闫　婷，孙春梅
（江苏大学 计算机科学与通信工程学院，江苏 镇江 212013）

摘　要：针对基于图的半监督学习方法在多媒体研究应用中忽略视频相关性的问题，提出了一种基于相关核映射线性近邻传播的视频标注算法。该算法首先通过核函数按照半监督学习调整后的距离计算出迭代标记传播系数；其次利用传播系数求得表示低层特征空间的样本，再根据视频相关性建模构造出语义概念间的关联表；最后完成近邻图的构造，并利用已标注视频信息迭代传播到未标注视频中，完成视频标注。实验结果表明，该算法不仅可以提高视频标注的准确度，还能弥补已标注视频数据数量的不足。

关键词：半监督学习；视频相关性；视频标注；语义相关性；线性近邻传播；核函数
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2013)02-0605-05
doi:10.3969/j.issn.1001-3695.2013.02.080

Video semantic annotation based on correlative kernel linear neighborhood propagation

ZHANG Jian-ming, YAN Ting, SUN Chun-mei
(*School of Computer Science & Telecommunication Engineering, Jiangsu University, Zhenjiang Jiangsu* 212013, *China*)

Abstract: In order to solve the problem that the graph-based semi-supervised learning methods neglect the video consistency in multimedia research area, this paper presented a new method video semantic annotation based on correlative kernel linear neighborhood propagation. The algorithm firstly structured the coefficient by kernel function, and through the coefficient to getting the samples that representative the low feature space. And then according to video correlation modeling, it structured the table between the semantic concepts. Finally, it completed the construction of graph, and used the video information that have been annotated spread to the video that not annotated. After that, the video annotation finished. The experimental results validate the effectiveness and superiority of the proposed method, the process of the video annotation addresses the insufficiency of labeled videos, improves the precision of annotation.

Key words: semi-supervised learning; video consistency; video annotation; semantic consistency; linear neighborhood propagation; kernel function

0　引言

随着存储设备和数字化设备的使用以及多媒体技术的发展，视频数据呈现几何级数增长的趋势。如何高效组织和检索这些视频数据成为了当前一个亟待解决的问题。在对视频的索引和检索中，最大的难题又在于如何缩小存在于视频底层特征和用户需求之间的语义鸿沟。

在对视频进行组织和检索时，常用的方法是从视频中提取出能对视频内容进行语义层次表述的元数据，然后再利用这些元数据对视频进行检索。视频语义标注（通常又称为视频概念检测、高层语义特征提取等）是获取这些元数据的一种基本方法。同时，基于内容的视频标注可以完成从视频底层特征到语义概念之间的映射，使得语义鸿沟被切分为两个更小的鸿沟[1]，即视频底层特征到语义概念之间的鸿沟和语义概念到用户需求之间的鸿沟。

视频标注实质是将多个相关的语义概念赋予到视频片段中，可分为基于人工的视频标注和基于机器学习的自动视频标注。完全使用人工标注是一项费时费力的工作，无法在大规模的数据集和概念集上应用。因此，使用机器学习方法来实现视频标注成为必然选择。

从机器学习的角度来说，视频标注的研究主要采用有监督学习、半监督学习和主动学习等方法。在很多实际应用中，获取有标记样本通常需要付出很大的代价。因此，半监督学习成为模式识别和机器学习中的重要研究领域。半监督学习[2]的主要思想是利用大量无标记数据自发地学习出数据的内在结构规律，再利用无标记数据与少量有标记数据之间的关系或相似度来进行指导，以期达到更好的学习效果。Yan 等人[3]分析了互训练在视频标注中的不足，并提出了一种改进的半监督交互特征方法；Wang 等人[4]使用半监督核密度估计的方法进行视频语义标注；He 等人[5]提出了用流形排序把已标注图像中的信息传播到未标注图像中；Yuan 等人[6]针对视频标注提出了一种基于特征选择的流形排序算法。

最近，Wang 等人[7]提到一种新的标签传递算法 LNP(linear neighborhood propagation)，它是一种基于图的半监督学习方

收稿日期：2012-07-08；**修回日期**：2012-08-18　　**基金项目**：国家自然科学基金资助项目(61170126)

作者简介：张建明(1964-)，男，江苏镇江人，教授，博士，主要研究方向为虚拟现实、图像处理、模式识别；闫婷(1988-)，女，山东济宁人，硕士研究生，主要研究方向为图像处理、模式识别(498567634@qq.com)；孙春梅(1987-)，女，山东济宁人，硕士研究生，主要研究方向为图像处理、模式识别.

法,使用在图上传播标记的方式来解决分类问题,但 LNP 算法不适合对底层特征空间比较复杂的视频进行标注。鉴于核技巧在模式识别领域取得的巨大成功,本文提出一种新的半监督学习算法,即基于相关核映射的线性近邻传播(correlative kernel linear neighborhood propagation,CKLNP)的算法,通过核函数把底层特征映射到一个非线性的特征空间中,解决 LNP 算法在视频标注中的限制,并将视频语义概念间的相互关系考虑到先验信息中。

1 LNP 算法简介

标记传播算法是基于图的半监督学习方法[8]之一,它使用在图上传播标记的形式来解决分类问题。文献[7]中提出的线性近邻传播(LNP)就是一种标记传播算法。一般来讲,所有的半监督学习算法都直接或间接地基于所谓的聚类假设:a)相邻的点可能属于一类;b)属于同一聚类或子流形结构的点属于同一类的可能性也很大。线性近邻传播继承局部线性嵌入的基本假设,即每个样本可以被它的近邻样本线性加权重建,并且样本的语义标号也可以被它的近邻样本的标号加权重建,且该重建的加权系数和样本重建的系数一样。公式描述如下:

$$x_i = \sum_{x_j \in N(x_i)} \alpha_j x_j \Rightarrow f_i = \sum_{x_j \in N(x_i)} \alpha_j f_j$$

其中:f_i 是 x_i 的语义标记,$N(x_i)$ 是样本 x_i 近邻样本。LNP 算法通过一系列重叠的线性近邻块来近似整个图 G,图中的边权重 W 是由标准二次规划问题求解得到的,然后合并边的权重形成对于 G 的权重矩阵。LNP 算法分为两步:a)近邻图的构造;b)传播已有标签数据的标号到剩余没有标签的数据上。

1.1 近邻图的构造

假设 $X = \{x_1, x_2, \cdots, x_l, x_{l+1}, \cdots, x_n\}$ 表示 R^d 空间中的 n 个数据对象,$L = \{1, -1\}$ 是标签集合,前 l 个数据点 $x_i \in X (1 \leqslant i \leqslant l)$ 表示已被标注的数据,对应的标记 $L_i \in L$,后 $n - l$ 个点表示未被标注的数据。X 所对应的图 $G = (V, E)$,顶点集合 $V = X$,E 表示边的集合,对于数据点 x_i 和 x_j 之间的关系,边的权重为 $w_{i,j}$。

LNP 算法[7]采用每一个数据点的近邻信息来构造整个图 G。为了方便,假设每一个数据点都能由它的近邻进行线性构造,并且能达到最优。因此,优化的目标是

$$\varepsilon = \min \sum_i \| x_i - \sum_{j:x_j \in N(x_i)}^{n} w_{ij} x_j \|^2$$

其中:x_j 是 x_i 的第 j 个近邻点;$N(x_i)$ 表示 x_i 的 k 个近邻点的集合;w_{ij} 表示 x_j 对 x_i 的贡献。加入两个约束:$\sum_{j:x_j \in N(x_i)}^{n} w_{ij} = 1$,$w_{ij} \geqslant 0$。可以很明显看出,$x_j$ 与 x_i 相似度越大,w_{ij} 的值就越大。因此,x_j 与 x_i 的相似性由 w_{ij} 来度量。可知 $w_{i,j} \neq w_{j,i}$,数据点 x_i 的 Gram 的矩阵 G^i 定义为 $G^i_{jk} = (x_i - x_j)^{\mathrm{T}} (x_i - x_k)$,进一步推导如下:

$$\varepsilon_i = \| x_i - \sum_{j:x_j \in N(x_i)}^{n} w_{ij} x_j \|^2 = \| \sum_{j:x_j \in N(x_i)}^{n} w_{ij} (x_i - x_j) \|^2 =$$

$$\sum_{j,k:x_j,x_k \in N(x_i)}^{n} w_{ij} w_{ik} (x_i - x_j)^{\mathrm{T}} (x_i - x_k) = \sum_{j,k:x_j,x_k \in N(x_i)}^{n} w_{ij} G^i_{jk} w_{ik}$$

每一个数据点的重构权重可由标注二次规划问题求解得到

$$\min w_{ij} \sum_{j,k:x_j,x_k \in N(x_i)}^{n} w_{ij} G^i_{jk} w_{ik}$$

$$\text{s.t.} \sum_{j:x_j \in N(x_i)}^{n} w_{ij} = 1 \quad w_{ij} \geqslant 0$$

求解完 n 个这样的二次规划问题,就可以构造出权重矩阵 W。权重矩阵的初始化为

$$W_{ij} = \begin{cases} w_{ij} & x_j \in N(x_i) \\ 0 & x_j \notin N(x_i) \end{cases}$$

1.2 基于 LNP 算法的视频数据标注

假设 F 表示定义在 X 上的分类函数集,$\forall f \in F$ 能对 x_i 指定 $f_i = f(x_i)$ 的实际值,而未被标注点 x_u 的类标记号可以由 $f_u = f(x_u)$ 来确定。每一次迭代,数据点一部分从它的近邻点中“吸收”标记信息,另一部分从它的初始标记信息中强化。这样在 $t+1$ 轮之后,x_i 的标记为

$$f_i^{t+1} = \alpha \sum_{j:x_j \in N(x_i)} w_{ij} f_i^t + (1 - \alpha) y_i$$

其中:$\alpha (0 < \alpha < 1)$ 是 x_i 从其近邻中接收的标记信息的权值;$y = (y_1, y_2, \cdots, y_n)^{\mathrm{T}}$,$y_i = L_i (i \leqslant l)$,$y_u = 0 (l + 1 \leqslant u \leqslant n)$,$f^{\mathrm{T}} = (f_1^t, f_2^t, \cdots, f_n^t)^{\mathrm{T}}$ 是预测的标记序列,初始化 $f^0 = y$。因此,可以将迭代方程改写为

$$f^{t+1} = \alpha w f^t + (1 - \alpha) y$$

可以用此公式更新每个数据点的标记信息,直到稳定状态。

1.3 LNP 算法的不足

1.2 节中已经提到 LNP 算法[7]是基于一个基本的假设:样本的标号可以由其近邻样本的标号加权得到,且公式表示为

$$x_i = \sum_{x_j \in N(x_i)} \alpha_j x_j \Rightarrow f_i = \sum_{x_j \in N(x_i)} \alpha_j f_j$$

定义从特征到语义的映射为 $f: x_i \rightarrow f_i$,可以得到

$$f_i = f(x_i) = f(\sum_{x_j \in N(x_i)} \alpha_j x_j)$$

$$f_i = \sum_{x_j \in N(x_i)} \alpha_j f_j = \sum_{x_j \in N(x_i)} \alpha_j f(x_j)$$

从而得到

$$f(\sum_{x_j \in N(x_i)} \alpha_j x_j) = \sum_{x_j \in N(x_i)} \alpha_j f(x_j)$$

上式说明了在特征空间的局部区域内从特征到语义标记的映射是线性的。在很多情况下(如在局部线性嵌入(LLE)[9]中),这种用来重建特征向量的局部线性假设是可行的,但在基于图的半监督学习中,在构造迭代标记传播系数时这种线性假设的表示能力表现出局限性。因此需要把线性传播扩展到非线性传播,以增强视频标注模型的表示能力。此外,从 LLE[9]的角度分析,LNP[7]假设语义分布是嵌入在特征空间中的一个一维流形,很显然一维流形不足以表示复杂的语义分布。同时为保证 LNP 的标记传播的收敛性,必须要求重建系数非负,这导致重建线性语义标记不理想。当语义分布非常复杂时,重建标记所使用的重建系数非负的语义局部线性假设的局限性更为突出。因此这个方法不能很好地解决视频语义标注问题,因为大部分视频数据的语义分布都非常复杂,也即 LNP 在语义分布局部线性假设上存在不足。

2 基于相关核映射线性近邻传播的视频语义标注算法

为解决上文所述的 LNP[7]算法的限制,使得视频标注模型在视频数据语义分布复杂的情况下也能表现出良好的性能。本文引入核函数,提出了 CKLNP 算法,该算法是对 LNP 算的改进。在 LNP 算法中,迭代标记传播系数是线性表示的,而非

线性表示的迭代标记传播系数更能增强标注模型能力。为将线性传播扩展到非线性传播上,构造出更准确的迭代标记传播系数,CKLNP 算法把底层特征通过核映射到一个高维的非线性特征空间中,由此重建的系数可以在这个非线性空间中计算。CKLNP 算法同样假设每个样本的标记可以由其邻居样本的标记线性重建,但是因为引入了核函数,将线性传播扩展到了非线性传播上,在核映射的空间中能有更合适的重建系数,所以 CKLNP 算法的线性重建要更为合理。

2.1　符号说明

假设 $X=(x_1,x_2,\cdots,x_l,x_{l+1},\cdots,x_n)$ 表示 R^d 空间中的 n 个数据对象;$C=(c_1,c_2,\cdots,c_b)$ 表示 b 个语义概念;函数 $f(x_i,c_p):X\times C\to R$ 表示样本 x_i 和语义概念 c_p 之间的函数关系,$\boldsymbol{f}$ 为大小为 $n\times b$ 的向量。为了讨论的方便,假设向量 $\boldsymbol{f}$ 的前 N_l 个元素已经被标注,且标注结果由向量 $\boldsymbol{y}$ 来表示,$\boldsymbol{f}$ 中剩余的 $N_u=n\times b-N_l$ 个元素是待标注的数据对象,标注结果由矩阵 $\boldsymbol{f}_u$ 来表示。利用矩阵 $\boldsymbol{S}^X=\lfloor S^X_{i,j}\rfloor_{n\times n}$ 表示在低层特征空间中的样本,$\boldsymbol{S}^C=[S^C_{p,q}]_{b\times b}$ 表示语义概念间的相关性。

2.2　通过核映射调整距离,得出表示低层特征空间的样本

下式是一个从输入特征空间 X 到映射空间 $\boldsymbol{\Phi}$ 的核映射:

$$\phi:X\to\boldsymbol{\Phi},x\to\phi(x)$$

数据被映射成 $\{\phi(x_1),\phi(x_2),\cdots,\phi(x_l),\phi(x_{l+1}),\cdots,\phi(x_n)\}$,本文选用径向基函数(RBF)作为核函数。点积的核矩阵 $\boldsymbol{K}$ 可以被表示为 $\boldsymbol{K}=(k_{i,j}),1\leqslant i\leqslant n,1\leqslant j\leqslant n$。其中,$k_{i,j}=\phi^{\mathrm{T}}(x_i)\phi(x_i)$,样本 $\phi(x_i)$ 的 k 个邻居样本集合 $N(\phi_i)$ 可以通过计算下式得到

$$\mathrm{dist}(\phi_i,\phi_j)=\|\phi(x_i)-\phi(x_j)\|=\sqrt{\phi_i^{\mathrm{T}}\phi_i-2\phi_i^{\mathrm{T}}\phi_j+\phi_j^{\mathrm{T}}\phi_j}=\sqrt{k_{ii}-2k_{ij}+k_{jj}}\tag{1}$$

从式(1)中可以观察到,距离不需要映射函数 $\phi(\cdot)$ 的形式,可以直接从核矩阵的计算中得到。

在计算距离时,可以将已标注样本的正例点、负例点以及未标注的样本加入到先验知识中,在不改变拓扑结构的前提下,尽量增大信息源的数据量可以得到更好的标注结果。本文将监督学习思想引入,调整图像点之间的距离,使用调整后的距离来进行非线性重构。

为增大原数据的信息量,可以借鉴监督式局部线性嵌入算法,构建 k 近邻的距离表示方法[10]:

$$D(x_i,x_j)=\begin{cases}\sqrt{1-\mathrm{e}^{\frac{-d^2(x_i,x_j)}{\beta}}}\\ \sqrt{\mathrm{e}^{\frac{-d^2(x_i,x_j)}{\beta}}}\end{cases}$$

上式分别表示 x_i 和 x_j 有相同标注和不同标注的情况。其中 $d(x_i,x_j)$ 表示 x_i 和 x_j 之间的欧式距离。基于这种想法,引入先验信息进行半监督学习,进一步改进距离求解的过程。当未知两图像点 x_i 和 x_j 是否同类,但 x_i 和 x_j 互为 k 近邻时,它们之间的距离大于它们同类时的距离,而小于它们之间关系为其他情况,这样公式[10]调整如下:

$$D(x_i,x_j)=\begin{cases}\sqrt{1-\mathrm{e}^{\frac{-d^2(x_i,x_j)}{\beta}}} & \text{有相同的标注}\\ \sqrt{1-\mathrm{e}^{-(\frac{d(x_i,x_j)}{\beta})^2}}+\sqrt{\frac{1}{2}\mathrm{e}^{(\frac{d(x_i,x_j)}{\beta})^2}} & \text{标注未知}\\ \sqrt{\mathrm{e}^{\frac{d^2(x_i,x_j)}{\beta}}} & \text{其他}\end{cases}$$

由此,通过核映射调整后的距离可以表示为

$$D(\phi_i,\phi_j)=\begin{cases}\sqrt{1-\mathrm{e}^{\frac{-\mathrm{dist}^2(\phi_i,\phi_j)}{\beta}}} & \text{有相同的标注}\\ \sqrt{1-\mathrm{e}^{-(\frac{\mathrm{dist}(\phi_i,\phi_j)}{\beta})^2}}+\sqrt{\frac{1}{2}\mathrm{e}^{(\frac{\mathrm{dist}(\phi_i,\phi_j)}{\beta})^2}} & \text{标注未知}\\ \sqrt{\mathrm{e}^{\frac{\mathrm{dist}^2(\phi_i,\phi_j)}{\beta}}} & \text{其他}\end{cases}\tag{2}$$

此时,可以运用 $D(\phi_i,\phi_j)$ 重新计算出样本的 k 个近邻点,并计算重建系数。在 CKLNP 算法中同样假设每个样本的标记可以由其近邻样本的标记线性重建。为了保证迭代的收敛性,加入两个约束条件到计算中,即 $w_{ip}\geqslant 0$ 和 $\sum_p w_{ip}=1$,这样可以通过求下面的问题来求得 ϕ_i 的最优重建系数:

$$\min\|\phi_i-\sum_{\phi_p\in N(\phi_i)}w_{ip}\phi_p\|^2\tag{3}$$

在计算完 w_{ip} 后,由此可以构造表示低层特征空间的样本的 $\boldsymbol{S}^X$:

$$\boldsymbol{S}^X=\begin{cases}w_{ip} & \phi_p\in N(\phi_i)\\ 0 & \text{else}\end{cases}\tag{4}$$

2.3　基于视频语义概念间的相关性建模

在完成了表示低层特征空间样本的 $\boldsymbol{S}^X$ 后,还需要得到表示样本语义概念间关联性的 $\boldsymbol{S}^C$。

如今已有很多种不同的半监督学习算法[11]被应用到多媒体的研究领域中,在这些方法中,都考虑了不同语义概念之间的相关性。语义概念间的相关性是指两个或多个概念有很大的可能性存在于同一图像中。将其考虑到视频标注中,虽然可以提高视频标注的准确率,但有时结果不精确。另外,即使彼此相关性很强的两个语义概念也并非就能提高视频标注的准确率,如语义概念“人脸”和“人”的相关性很强,它们几乎同时出现,但它们对标注结果意义不大。其原因是有相关噪声的存在,且在这种标注模型中并没用考虑到语义概念间的相关性是有向的。例如在一幅图像中,如果检测到语义概念“沙漠”,实例证明检测到“天空”的概率比较大;但反之,“天空”的出现不一定伴随着“沙漠”。由此可以知道,语义概念“沙漠”对检测到语义概念“天空”是有效的;相反,“天空”对检测到“沙漠”没有很大的作用。从此例中可以得出,在视频语义概念标注模型中,方向性是一个至关重要的因素。为了解决上述问题,CKLNP 算法不仅归一化交互信息,同时构建一个相关表来抑制相关噪声的出现,同时保证了相关性有向。

为了抑制相关噪声的出现,利用专业知识构造一个二进制相关表 $K(K(m,n)\in\{0,1\})$,从相关表中可以得出一个概念对另外一个概念的标注是否有用。在构建相关表时,先邀请五个没有专业知识的志愿者,让他们表示两个概念间是否有相关性,然后将其投票结果作为两个概念间是否有相关性的最终结果。例如,如果志愿者认为语义概念 m 对检测语义概念 n 是有用的,则 K 的值为 1,否则为 0。在志愿者确定相关表时,有一些基本的规则:a)如果当前的语义概念 m 有很多的正例,而语义概念 n 几乎同时伴随着 m 出现,则定义 n 对检测到 m 没用,如上文中提到的“人脸”与“人”的例子;b)如果语义概念 m 有很少的正例,语义概念 n 有时会伴随 m 出现,同时语义概念 n 并不与很多其他的语义概念同时出现,则定义语义概念 n 对检测到语义概念 m 有用;c)如果语义概念 m 与很多其他的语义概念伴随出现,则定义语义概念 m 对检测到其他的语义概

念没用。例如,在很多实例中,语义概念“汽车”和“飞机”经常伴随出现,而“飞机”几乎不和除了“天空”的其他语义概念伴随出现,而“天空”又几乎和很多其他的语义概念伴随出现,“汽车”和“飞机”有着相似的视觉感知效果。可以样认为,“汽车”对检测到“飞机”有用,而“天空”对检测到“飞机”没用的相关表是通过志愿者的学习得到的,虽然不是很精确,但是本文的目的只是想确定一个语义概念对检测到另一个语义概念是否有用,所以相关表仍然可取,并且相关表可以扩展。如果新添加一个语义概念,只需给相关表再添加一行一列,并不需要重新构造。志愿者只需要 30 min 就可以确定一个相关表,这与标注视频相比节约了大量的人力、物力、时间,提高了视频标注的效率。

在文献[12]中语义概念 m 的边际熵定义为

$$H(m) = -\sum_{y_m \in \{+1,-1\}} p(y_m)\log_{10}^{p(y_m)}$$

语义概念 m 和语义概念 n 的互信息[12]定义为

$$M(m,n) = \sum_{y_m,y_n} p(y_m,y_n)\log\frac{p(y_m,y_n)}{p(y_m)p(y_n)}$$

令 $Q = \min\{H(m),H(n)\}$,则语义概念 m 和语义概念 n 的标准化交互信息可以定义为

$$M^*(m,n) = \frac{M(m,n)}{Q} \tag{5}$$

标记先验 $p(y_m)$、$p(y_n)$可以从训练数据集中估计到。显然,语义概念 m 和语义概念 n 之间的关联性越强,$M^*(m,n)$的值越大。

通过以上分析,可以得出语义概念 m 和语义概念 n 之间相关性定义式如下:

$$C(m,n) = K(m,n)M^*(m,n) \tag{6}$$

值得注意的是:因为 $K(m,n) \neq K(n,m)$,使得 $C(m,n) \neq C(n,m)$,这样也就能保证语义概念之间相关性程度是有向的。令 $S^C = C$,则 S^C 可以求出。

2.4 利用相关性和低层样本构造近邻图并进行视频标注

根据基于图的半监督学习理论[13],可以最小化 $E = f^T L^{X,C} f$ 以获得最优标注结果。其中:

$$L^{X,C} = D^{X,C} - S^{X,C}S^{X,C} = S^X \otimes S^C \tag{7}$$

其中:⊗代表矩阵间的乘积;$D^{X,C}$是一个对角矩阵,且对角线上的元素定义为 $D^{X,C}_{(i,p)(i,p)} = D^X_i D^C_p$,$D^X_i = \sum_{j=1}^{n} S^X_{i,j}$,$D^C_p = \sum_{q=1}^{b} S^C_{p,q}$。对 $S^{X,C}$标准化得到

$$\hat{S}^{X,C} = (D^{X,C})^{-1} S^{X,C} \tag{8}$$

分割 $\hat{S}^{X,C}$得到

$$\hat{S}^{X,C} = \begin{bmatrix} \hat{S}^{X,C}_{l,l}, \hat{S}^{X,C}_{l,u} \\ \hat{S}^{X,C}_{u,l}, \hat{S}^{X,C}_{u,u} \end{bmatrix}$$

其中:$\hat{S}^{X,C}_{i,j} \geqslant 0$,$\sum_{j=1} \hat{S}^{X,C}_{i,j} = 1$,且 $\hat{S}^{X,C}_{u,u}$与 $\hat{S}^{X,C}_{u,l}$的谱半径均小于 1。

为维持已标注数据对象的标记强度,令 $f_l \equiv y$,使得最初的标记在传播过程中不会被淡化,可以令 $f = \begin{pmatrix} f_l \\ f_u \end{pmatrix}$。事实上,相关核映射线性近邻传播算法本质上就是得到 f_u,因此标注可以通过迭代

$$f_u^{(t+1)} = \hat{S}^{X,C}_{u,l} f_l^{(t)} + \hat{S}^{X,C}_{u,u} f_u^{(t)} \tag{9}$$

标记传播完成。其中 $f_l^{(t)} \equiv y$,迭代直到收敛,这实际上是一个从邻域样本标记到当前样本标记的信息传播过程。

2.5 CKLNP 算法步骤

本文受线性近邻传播算法[7]、核技巧以及半监督学习[11]方法的启发,在此基础上提出基于相关核映射线性近邻传播的视频标注算法。该算法满足半监督学习的两个先验假设[7]:a)平滑性假设,具有相似特征的数据点具有一致的类标签;b)适应性假设,具有同一结构的数据点具有一致的类标签。针对上文中的叙述,给出本文提出的基于相关核映射线性近邻传播的视频标注算法的基本步骤:

a)核利用径向基函数,计算对应于 X 的核矩阵 $K = (k_{ij})$,$1 \leqslant i,j \leqslant n$。

b)使用式(2)计算 Φ 中的每个样本 ϕ_i 的 k 个最近邻样本。

c)根据式(3)求出迭代标记系数,并通过式(4)构造低层特征空间中的样本矩阵 S^X。

d)从训练集中计算出各个语义概念出现的概率,由此计算出语义概念间的标准化交互信息。

e)构造相关表 K,根据表 K 计算 C,从中构造出代表语义概念相关性的矩阵 S^C。

f)根据式(7)求解 $S^{X,C}$,并对 $S^{X,C}$根据式(8)规范化得到 $\hat{S}^{X,C}$;迭代式(9)直到收敛,这样就能得到未标记样本的实数值语义标记。

该算法的核心思想是让每个已标注的顶点迭代地传播其标注信息到未标注的顶点,直到构造出的概率图达到完全稳定状态。

3 实验结果

3.1 实验样本

本文设计了实验来验证提出的相关核映射线性近邻传播算法的性能,实验所用样本来自标准视频数据集 TRECVID2007 中的视频片段。首先为视频片段划分镜头;然后为每个镜头提取关键帧,形成实验样本集;最后从样本集中挑选出样本组成训练集,其余样本组成测试集。

在提取镜头的关键帧后,选取关键帧的图像视觉特征作为镜头的低层特征,具体如下:

a)HSV 颜色特征。根据人类对色调(H)、饱和度(S)、纯度(V)感知能力不同,按 $8 \times 3 \times 3$ 进行非等量量化,即 H 被分成 8 个等级,S 被分成 3 个等级,V 被分成 3 个等级。

b)纹理特征。使用 Tamura 纹理特征中的粗糙度、对比度和方向度这三个特征。

c)形状特征。采用图像的 Hu 不变矩。

本文所选用的语义概念包括多个不同的类型,分别为 person、meeting、car、sports、building、weather、road、animal、mountain、outdoor。从与各语义概念相关的镜头中提取部分关键帧,如图 1 所示。

3.2 实验结果与分析

如何对标注性能进行合理评价,这是设计实现一个标注系统过程中的一个关键问题,也是本文进行实验目的所在。本文系统性能由 TRECVID 任务中官方性能度量平均查准率(average precision,AP)和平均标全率(average recall,AR)来评估。前者表示标注的准确程度,反映了系统减少噪声的能力;

后者表示标注是否全面,反映了系统是否漏检的能力。把 APs 和 ARs 在十个概念上平均就可以得到 mean average precision (MAP)和 mean average recall(MAR),这是最终的评价度量。计算方法[14]如下:

$$\mathrm{MAP}=\frac{1}{k}\sum_{i=1}^{k}\mathrm{precision}[c_i],\mathrm{MAR}=\frac{1}{k}\sum_{i=1}^{k}\mathrm{recall}[c_i]$$

其中:k 为语义的总数,c_i 表示第 i 个语义。precision$[c_i]$、recall$[c_i]$计算公式为

$$\mathrm{precision}[c_i]=\frac{N_{\mathrm{correct}}[c_i]}{N_{\mathrm{pLabel}}[c_i]}$$

$$\mathrm{recall}[c_i]=\frac{N_{\mathrm{correct}}[c_i]}{N_{\mathrm{label}}[c_i]}$$

其中:$N_{\mathrm{correct}}[c_i]$表示用 c_i 正确标注测试集中视频镜头的数目,$N_{\mathrm{pLabel}}[c_i]$表示用 c_i 标注测试集中视频镜头的数目,$N_{\mathrm{label}}[c_i]$表示测试集中与 c_i 相关的实际视频镜头的数目。

图 1　十个语义概念的部分关键帧

本文实验是在安装了 Windows XP 的微机上进行的,设计了在标准视频数据库集 TRECVID 2007 中的实验来验证基于相关核映射线性近邻传播算法的性能。实验中使用 K-NN 来寻找近邻点,但搜寻近邻节点这一过程消耗的时间比较长,而这又是基于图的半监督学习算法必不可少的一部分。实验时,先取小部分的样本进行实验,通过设置不同的 k 值,计算实验消耗的时间,以得出 k 的最优值。根据结论可得出,一般 k 为 30 时最优。

为评估视频标注性能,将 CKLNP 算法与两种广泛使用的半监督学习方法(SVM[15]和 LNP[7])进行比较。

若以 AP 为衡量标准,其结果如图 2 所示。从图 2 中可以看出,以 AP 为衡量标准时,与 LNP 算法相比,CKLNP 算法在检测大部分的语义概念时都表现出良好的性能,除了在语义概念 weather 上略有不足,其主要原因在于语义概念 weather 比较难以检测到,并且实验结果比较随机。在语义概念 person、meeting、car、sports、weather、road、animal、outdoor 上,CKLNP 算法获得的 AP 值要比通过 SVM 算法获得的高。通过以上数据分析,CKLNP 算法从整体上优于其他两种算法,可以得出 CKLNP 算法提高了查准率。

若以 AR 为衡量标准时,三种算法结果如图 3 所示。从图 3 中可以看出,CKLNP 算法在检测语义概念 meeting、building、road、mountain 时得到的 AR 值与经典的 SVM 算法相比近似相等,而在其他的语义概念上获得的 AR 值都比通过 SVM 算法[15]获得的高;而 LNP 算法[8]获得的 AR 值都比 CKLNP 算法获得的 AR 值低一些。因此 CKLNP 算法在整体范围上要优于 LNP 和 SVM 算法,这也证明了 CKLNP 算法在提高视频标注标全率上是有益的。

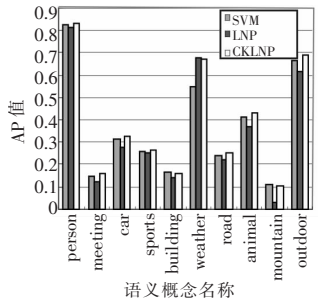


图 2　三种算法关于 AP 的比较结果

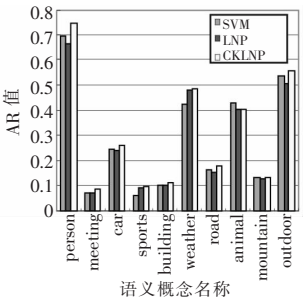


图 3　三种算法关于 AR 的比较结果

为评估视频标注的准确度,将 CKLNP 算法与其他两种算法(SVM 和 LNP)进行比较。

若以 MAP 为衡量标准,其结果如表 1 所示。

表 1　三种算法的 MAP 值比较

method	MAP	improvement/%
CKLNP	0.389 774	–
SVM	0.369 118	5.6
LNP	0.352 659	10.5

从表 1 可以看出,CKLNP 算法的 MAP 值为0.389 774,SVM 算法的 MAP 值为 0.369 118,按照 MAP 值作为评估标准,CKLNP 相对于 SVM 算法提高了 5.6%;LNP 算法获得的 MAP 值为 0.352 659,CKLNP 算法相对于 LNP 算法提高了 10.5%。从这组数据中可以看出,本文提出的 CKLNP 算法在减少噪声方面有所改善,并且提高了视频标注的准确程度。

若以 MAR 为衡量标准,其结果如表 2 所示。

表 2　三种算法的 MAR 值比较

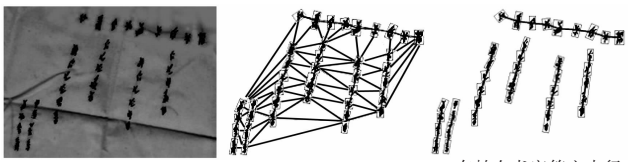
method	MAR	improvement/%
CKLNP	0.304 732	–
SVM	0.285 574	6.7
LNP	0.284 309	7.2

从表 2 可以看出,CKLNP 算法获得的 MAR 值为 0.304 732,LNP 算法的 MAR 值为0.284 309,SVM 算法的 MAR 值为 0.285 574。由此可以得出,按照 MAR 作为评估标准,CKLNP 算法相对 LNP、SVM 算法分别提高了 7.2% 和 6.7%。这说明了 CKLNP 算法获得了更高的标全率,系统的性能有所提升。所有这些数据都说明了在视频语义标注中,相关核映射线性近邻传播的视频标注算法要比一般的半监督学习方法更有效。

4　结束语

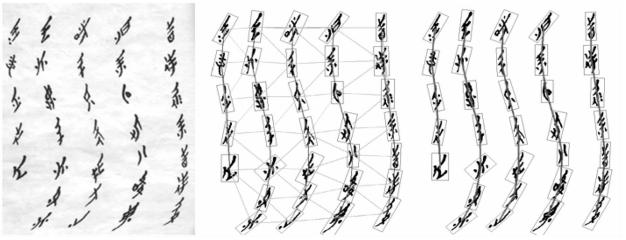
本文针对可以利用大量未标注样本来改善学习这一特点提出了一种新的基于图的半监督学习方法 CKLNP 来进行自动视频标注。该方法考虑到核技巧在模式识别邻域中的成功应用,并且将语义概念之间的关系融合到已有的 LNP 算法中。本文还构造了一个相关表用来抑制相关噪声的出现,并且保证了语义概念之间的方向性,这使得语义概念间的关联关系更为准确。实验结果表明,本文提出的 CKLNP 方法在视频标注应用中具有良好的性能。目前基于分类的张量学习方法引起了研究者的注意,将语义概念间的关联性考虑到张量学习中是一个不难解决的问题,所以未来的研究课题是如何将语义概念间的关联性融入到张量学习中。

(下转第 630 页)



(a) 布帕女书字符图像 (b) 建立链接模型 (c) 布帕女书字符文本行提取结果

图 7 布帕女书字符图像文本行提取结果



(a)弯曲女书字符图像 (b)建立链接模型 (c)弯曲女书文本行提取结果

图 8 弯曲女书字符图像文本行提取结果

实验结果表明,本文算法能够有效地提取出扇面、布帕等不同载体上女书字符图像中多方向的文本行结构以及少许弯曲的女书文本行结构。

3 结束语

本文实现了一种用带权重的三角网结构自动提取脱机手写女书多方向文本行的方法。实验结果表明,该方法能够有效提取出扇面、布帕等不同载体上女书字符图像中的多方向女书文本行,并且能够提取出布帕图像中存在弯曲的女书文本行,为女书字符切分以及识别奠定了基础。

参考文献:

[1] DOS SANTOS R P, CLEMENTE G S, REN T I, *et al*. Text line segmentation based on morphology and histogram projection[C]// Proc of the 10th International Conference on Document Analysis and Recognition. 2009:651-655.

[2] 李庆福. 女书文化研究[M]. 北京:人民出版社,2009.

[3] 肖人岳,秦慕婷. 一种复杂文本图像中快速文本行检测算法[J]. 科学技术与工程,2008,8(23):6253-6257.

[4] LIU Xiao-qing, SAMARABANDU J. Multiscale edge-based text extraction from complex images[C]//Proc of IEEE International Conference on Multimedia and Expo. 2006:1721-1724.

[5] 敖翔,戴国忠,王宏安. 基于感知的多方向在线手写笔迹文本行提取[J]. 计算机辅助设计与图形学学报,2007,19(1):14-19.

[6] PAL U, SINHA S, CHAUDHURI B B. Multi-oriented text lines detection and their skew estimation[C]//Proc of Indian Conference on Computer Vision, Graphics and Image Processing. 2002:270-275.

[7] OUWAYED N, BELAD A. Multi-oriented text line extraction from handwritten Arabic documents[C]//Proc of the 8th IAPR International Workshop on Document Analysis Systems. 2008:339-346.

[8] 周培德. 平面点集三角剖分的算法[J]. 计算机辅助设计与图形学学报,1996,8(4):259-264.

[9] 冈萨雷斯. 数字图像处理[M]. 阮秋琦,阮宇智,译. 2 版. 北京:电子工业出版社,2007.

[10] 周培德. 计算几何——算法分析与设计[M]. 北京:清华大学出版社,2000:61-62.

[11] PIRZADEH H. Computational geometry with the rotating calipers [D]. Montréal: School of Computer Science, McGill University, 1999.

[12] AO Xiang, LI Jun-feng, WANG Xu-gang, *et al*. Structuralizing digital ink for efficient selection[C]//Proc of the 11th International Conference on Intelligent User Interfaces. 2006:148- 154.

[13] 胡勇,王国胤,杨勇. 人脸特征约束点的三维表情合成[J]. 计算机应用研究,2012,29(2):754-756.

(上接第 609 页)

参考文献:

[1] HAUPTMANN A G. Lessons for the future from a decade of informedia video analysis research[C]//Proc of the 4th International Conference on Image and Video Retrieval. 2005:1-10.

[2] SONG Y, HUA Xian-sheng, DAI Li-rong, *et al*. Semi-automatic video annotation based on active learning with multiple complementary predictors[C]//Proc of Workshop on Multimedia Information Retrieval. 2005:97-104.

[3] YAN R, NAPHADE M. Semi-supervised cross feature learning for semantic concept detection in videos[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2005:657-663.

[4] SONG Y, HUA Xian-sheng, DAI Li-rong, *et al*. Semi-automatic video annotation based on active learning with multiple complementary predictors[C]//Proc of ACM International Workshop on Multimedia Information Retrieval. 2005:97-104.

[5] HE Jing-ni, LI Ming-jing, ZHNAG Hong-jiang, *et al*. Generalized manifold-ranking based image retrieval[J]. IEEE Trans on Image Processing, 2006, 15(10):3170-3177.

[6] YUAN Xun, HUA Xian-sheng, WANG Meng, *et al*. Manifold-ranking based video concept detection on large database and feature pool [C]//Proc of ACM Multimedia Conference. 2006.

[7] WANG Fei, WANG Jing-dong, ZHANG Chang-shui, *et al*. Semi-supervised classification using linear neighborhood propagation[C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2006:160-167.

[8] 卢汉清, 刘静. 基于图学习的自动图像标注[J]. 计算机学报, 2008, 31(9):1630-1639.

[9] SAUL L K, ROWEIS S T. Think globally, fit locally: unsupervised learning of low dimensional manifolds [J]. Journal of Machine Learning Research, 2003, 4:119-155.

[10] GENG Xin, ZHAN De-chuan, ZHOU Zhi-hua. Supervised nonlinear dimensionality reduction for visualization and classification[J]. IEEE Trans on Systems, Man, and Cybernetics, Part B: Cybernetics, 2005, 35(6):1098-1107.

[11] YANG Gao-ming, YANG Jing, ZHANG Jian-pei. Semi-supervised clustering-based anonymous data publishing [J]. Journal of Harbin Engineering University, 2011, 32(11):1489-1494.

[12] QI Guo-jun, HUA Xian-sheng, RUI Yong, *et al*. Correlative multi-label video annotation[C]//Proc of ACM Multimedia Conference. 2007: 17-26.

[13] ZHA Zheng-jun, MEI Tao, WANG Jing-dong, *et al*. Graph-based semi-supervised learning with multi-label[C]//Proc of IEEE International Conference on Multimedia and Expo. 2008:1321-1324.

[14] TANG J, HUA Xian-sheng, MEI Tao, *et al*. Video annotation based on temporally consistent Gaussian random field [J]. Electron Lett, 2007, 43(8):448-449.

[15] TONG S, CHANG S F. Support vector machine active learning for image retrieval[C]//Proc of the 9th ACM International Conference on Multimedia. New York: ACM Press, 2001:107-118.