视频语义分析两级多模式融合算法

魏 维[1] 李千目[1] 刘凤玉[1] 许满武[2]

[1](南京理工大学计算机科学与技术系，南京 210094) [2](南京大学计算机科学与技术系，南京 210008)

摘 要 为了全面准确地获取视频高层语义信息，提出了一种基于仿生的视频语义分析两级多模式融合算法。该算法仿照人脑多感觉融合机理，先将视频中多模式特征按不同类别划分为组，然后对每一组中的多模低层特征用层次隐马尔可夫模型(HHMM)进行数据融合；同时将以似然率表示的多个低层融合结果作为高层融合的输入，再通过基于核的非线性算法把输入空间变换到高维特征空间；最后在特征空间中求取最优线性分类面，即可得到最终的多模式两级融合结果。实验表明，该方法不仅能有效融合视频中的多模式特征，而且能获取全面、准确的高层语义信息。

关键词 多模式融合 视频语义概念 多层次分析 决策融合

中图法分类号：TP391.41 TP18 **文献标识码**：A **文章编号**：1006-8961(2007)05-0893-06

Two Level Multimodal Fusion Algorithm for Semantic Video Analysis

WEI Wei[1], LI Qian-mu[1], LIU Feng-yu[1], XU Man-wu[2]

[1](*Department of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing* 210094)

[2](*Department of Computer Science and Technology, Nanjing University, Nanjing* 210008)

Abstract To extract video semantic concepts combing different modalities, a two level multimodal fusion method for video semantic concept analysis is proposed. Multimodal features of video are divided into several groups. The fact that each group of has distinct features, a hierarchical hidden Markov models(HHMM) is constructed for the purpose of first-level fusion. Then outputs of first-level fusion are combined using a kernel function, by which a hyper-plane with better classification for video semantic concept is obtained. The results of experiments comparing to other fusion methods support that the two-level fusion method utilizes different modal feature in semantic concept analysis, and could effectively combine multimodal features.

Keywords multimodal fusion, video semantic concept, multilayer analysis, decision fusion

1 引 言

视频语义分析主要是研究特征描述与高层语义概念间的关系，其最终目的是从各种视频特征及相关的原始视频数据中自动提取视频语义概念。视频是多模式特征的混合媒体，但单一模式的特征只能反映局部、侧面的信息。因此如何有效地融合多种模式特征，以获取全面、准确的高层语义信息，已成为视频分析领域中的一项重要的研究内容。

针对视频语义概念分析，本文提出了一种两级的多模式视频数据融合算法。此算法把视频的多种模特征划分为多组，并形成两层(级)结构。同一组的多个低层特征采用 HHMM (hierarchical hidden Markov models)进行数据融合，而各低层融合的结果则采用基于核的非线性方法进行二次融合。

2 模式划分与融合原理

2.1 人脑多模式知觉融合

由人体生理学知，在大脑两半球的皮层上存在着一些初级专门区域，各种感觉器官(眼、耳、鼻等)

基金项目：国家自然科学基金项目(60273035)；江苏省科技攻关项目(BE2003064)

收稿日期：2005-10-17；改回日期：2006-02-28

第一作者简介：魏维(1976～)，男。2003 年获南京理工大学硕士学位，现为南京理工大学计算机系博士研究生。主要研究方向为视频内容分析、语义视频检索。E-mail：weiwei863@hotmail.com

受到刺激后由中枢神经传递给相应的区域,然后在这些区域中具有相同感受野的多种特征检测细胞聚集在一起,实现同一种感觉模式中各种刺激(信息)的综合反应,就形成简单的知觉;而联络区皮层的多模式感知细胞,则将多种模式的感觉信息综合为复杂的知觉[1]。人脑的这种多模式知觉融合具有特有的两层结构。

图 1 为人在爆炸现场形成“产生爆炸”高层语义的过程。其过程是:①视觉分析器将光信号转变成生物电信号,并通过神经系统传至大脑,在结合经验、记忆的基础上,进行分析、判断、识别等极为复杂的过程后,在大脑中形成物体的形状、颜色等概念;②利用形状、明暗及颜色的变化,在视觉分析器与运动分析器(眼肌活动等)的协调作用下,融合各视觉信息(大脑皮层上的视觉区域)产生了“火焰”和“浓烟”视觉的语义;③听觉器官将音强、音调等声音特征在相应区域融合得出“隆隆”爆炸声语义;④在大脑皮层嗅觉区域,空气中的各种气味刺激融合后得到“焦糊味”的嗅觉语义;⑤皮肤把温度和压力等信息收集送到知觉区域,经过信息融合得到“热浪袭来”的感觉语义;⑥最后多模式感知细胞融合各种初级感觉语义就得到“产生爆炸”的语义。

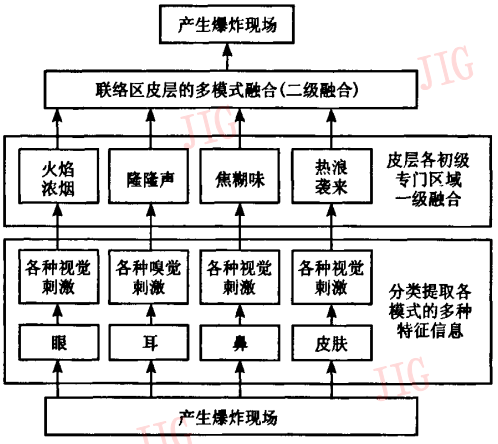

图 1 人脑多模式知觉融合结构

Fig. 1 Multimodal information fusion structure of brain

2.2 仿生融合原理

2.2.1 模式划分

视频中每种模式特征所涵盖的语义信息都很丰富,且在不同的应用中具有相互独立又相互补充的特点。各种模式的特征都是信息载体的一种形式,但对视频数据模式的划分因人而异,还没有统一标准。如文献[2]中采用的多种模式为颜色直方图、运动强度、声音特征和可视概念矢量。而文献[3]中则将声音、人脸、分割的镜头分别作为单个模式。文献[4]用颜色和运动量两模式来分析足球比赛中“精彩镜头回放”。

本文将视频中的多模式特征统一划分为两层(级),第 1 层分为图像、声音、文本等模式组,其中每组高层模式中又包括多个同等级别的低层模式特征。图 2 是对视频数据进行两层划分的结果。

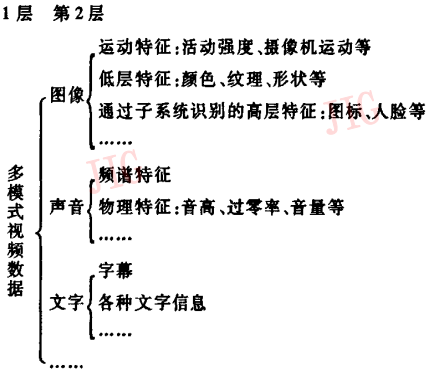

图 2 两层的多模式划分

Fig. 2 Two layer partition of multimodal feature

2.2.2 融合策略

现有的多模式融合算法主要是基于启发式策略的。其常用的模式融合又分为特征融合(feature fusion)和决策融合(decision fusion)两大类。其中特征融合(早期融合)指将多模特征同时作为模型的输入来处理,但特征融合常产生高维特征矢量,从而导致维数的急剧升高,而且除维数问题外,在不同模式间存在异步和动态交换时,还存在特征融合性能较差的问题;决策融合(后期融合)指各模式特征分别按对应模型单独处理,最后把多个模式的识别结果进行融合。这种方法把不同模式所得识别结果在高层进行融合,并产生一个整体/全局的语义决策。

通常直接将低层特征映射到高层语义概念比较困难,但多层次分析可将高层语义分解为一系列可识别的低层原型(primitives)及各原型和高层语义间的关系。低层原型与低层特征可直接产生映射。这样从低层原型事物出发,通过推断便可提取语义概念。本文提出一种采用多模式融合和多层次分析相结合的策略,其中高层的融合是典型的决策融合类型。

2.2.3　两级融合算法原理

先将视频数据中多模特征按 2.1 节方法划分为两层结构,然后采用 HHMM 对同一组的低层模式特征进行低层次融合,而各组高层模式类间的融合再采用一级(高层)融合。这样整个数据融合过程就形成了特有的两级融合结构。视频数据两级融合算法原理如图 3 所示。

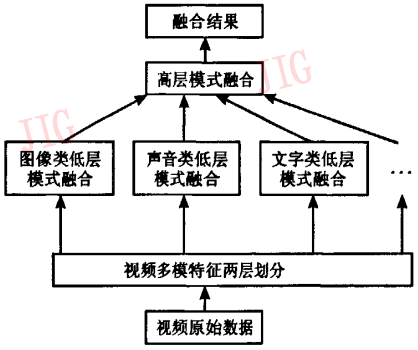

图 3　视频数据两级融合原理

Fig. 3　Principle of two level fusion for video data

本文解决高层融合的思路是先把低层融合的分类结果以似然率的形式表示(通常分类结果取与最大似然率对应的类别);然后将其作为高层融合的输入;同时把输入空间升维,使得视频语义概念在高维空间中线性可分。这样多类的低层融合的结果就得到进一步融合,并产生统一的分类融合预测函数 f。高层的融合是以基于核的统计学习为基础,通过支持向量的内积运算来达到进行高维空间变换的效果[5]。

3　融合模型与算法

3.1　低层融合模型

设 $\boldsymbol{\Sigma}$ 为有限的 alphabet 空间,HHMM 具有多层随机过程结构。设有限的可观察符号序列为 $\boldsymbol{O}=o_1o_2\cdots o_T$。HHMM 状态表示为 $q_i^{(d)}$ ($d\in\{1,\cdots,D\}$),其中,i 为状态序数,d 是层序数,产生状态的层次序数是 D。内部状态 $q_i^{(d)}$ 的子状态个数表示为 $|q_i^{(d)}|$。在不产生歧义的前提下,可省略状态数,用 $q^{(d)}$ 表示 d 层的一个状态。每一个内部状态对应一个状态转移矩阵 $\boldsymbol{A}^{q^{(d)}}=(a_{i,j}^{q^{(d)}})$,其中 $a_{i,j}^{q^{(d)}}=P(q_j^{(d+1)}\mid q_i^{(d+1)})$ 是从 i 状态 $q_i^{(d+1)}$ 到 j 状态 $q_j^{(d+1)}$ 的水平转移概率,两者都是 $q^{(d)}$ 的子状态。相似的,$\boldsymbol{\Pi}^{q^{(d)}}=\{\pi^{q^{(d)}}(q_i^{(d+1)})\}=\{P(q_i^{(d+1)}\mid q^{(d)})\}$ 表示 $q^{(d)}$ 到其子状态的初始概率分布。如果 $q^{(d+1)}$ 也是内部状态,那么 $\pi^{q^{(d)}}(q_i^{(d+1)})$ 也可解释为垂直转换概率。每个产生状态的唯一参量为其可观察符号的输出概率矢量 $\boldsymbol{B}^{q^{(D)}}=\{b^{q^{(D)}}\}$,即产生状态 $q^{(D)}$ 观察到符号 $\sigma_k\in\boldsymbol{\Sigma}$ 的概率。整个 HHMM 模型参数集表示为

$$\boldsymbol{\lambda}=\{\boldsymbol{\lambda}^{q^{(d)}}\}_{d\in\{1,\cdots,D\}}=\{\{\boldsymbol{A}^{q^{(d)}}\}_{d\in\{1,\cdots,D\}},\{\boldsymbol{\Pi}^{q^{(d)}}\}_{d\in\{1,\cdots,D\}},\{\boldsymbol{B}^{q^{(D)}}\}\} \tag{1}$$

模型参数估计:给定一个 HHMM 结构及一个或多个可观察序列 $\{\boldsymbol{O}_t\}$,找到模型最可能的参数 $\tilde{\boldsymbol{\lambda}}$,$\tilde{\boldsymbol{\lambda}}=\arg\max_{\boldsymbol{\lambda}}P(\{\boldsymbol{O}_t\}\mid\boldsymbol{\lambda})$。可用推广的向前-向后算法来估计 HHMM 的最大似然参数。参数迭代算式为[6]

$$\hat{\pi}^{q^{(1)}}(q_i^{(2)})=\chi(t,q_i^{(2)},q^{(1)}) \tag{2}$$

$$\hat{\pi}^{q^{(d-1)}}(q_i^{(d)})=\frac{\sum_{t=1}^{T}\chi(t,q_i^{(d)},q^{(d-1)})}{|q^{d-1}|\sum_{t=1}^{T}\sum_{t=1}^{T}\chi(t,q_i^{(d)},q^{(d-1)})}(2<d<D) \tag{3}$$

$$\hat{a}_{i,j}^{q^{(d-1)}}=\frac{\sum_{t=1}^{T}\xi(t,q_i^{(d)},q_j^{(d)},q^{(d-1)})}{|q^{(d-1)}|\sum_{t=1}^{T}\sum_{t=1}^{T}\xi(t,q_i^{(d)},q_j^{(d)},q^{(d-1)})}=\frac{\sum_{t=1}^{T}\xi(t,q_i^{(d)},q_j^{(d)},q^{(q-1)})}{\sum_{t=1}^{T}\gamma_{\text{out}}(t,q_i^{(d)},q^{(d-1)})} \tag{4}$$

$$\hat{b}_{q_i^{(D)}}^{q^{(D-1)}}(v_k)=\frac{\sum_{o_t=v_k}\chi(t,q_i^{(D)},q^{(D-1)})+\sum_{T>1,o_t=v_k}\gamma_{\text{in}}(t,q_i^{(D)},q^{(D-1)})}{\sum_{t=1}^{T}\gamma(t,q_i^{(D)},q^{(D-1)})+\sum_{t=2}^{T}\gamma_{\text{in}}(t,q_i^{(D)},q^{(D-1)})} \tag{5}$$

其中,$\xi(t,q_i^{(d)},q_j^{(d)},q^{(d-1)})$ 为在时刻 t,$q^{(d-1)}$ 的子状态从 $q_i^{(d)}$ 水平转移到 $q_j^{(d)}$ 的概率。$\gamma_{\text{in}}(t,q_i^{(d)},q^{(d-1)})$ 为在 o_t 产生前,水平转移到 $q_i^{(d)}$ 状态的概率。$\gamma_{\text{out}}(t,q_i^{(d)},q^{(d-1)})$ 是在 o_t 产生后,状态 $q_i^{(d)}$ 水平转移到其他状态的概率。$\chi(t,q_i^{(d)},q^{(d-1)})$ 表示在 t 时刻产生和选择活动状态 $q_i^{(d)}$ 的概率,v_k 表示产生状态时产生的可观察符号。用以上式(2)~式(5)进行迭代计算,其收敛的值即为所求参数。

3.2　高层基于核的非线性融合

高层的融合采用基于核的统计学习方法[7,8]。设 $\boldsymbol{S}$ 为输入空间,$n_{\text{dim}}=\dim(\boldsymbol{S})$。$\boldsymbol{X}$,$\boldsymbol{C}$ 分别为训练样本和视频语义概念类别标签。$\boldsymbol{S}_{\text{feature}}$ (fusion

space)是高层模式融合的特征空间。Φ 表示变换：$s \rightarrow s_{\text{feature}}$，则基于核的非线性变换算法如下：

$$\Phi: \mathbf{R}^{n_{\text{dim}}} \rightarrow s_{\text{feature}}$$

$$s \mapsto \Phi(s) \quad s_1, \cdots, s_n \in \mathbf{R}^{n_{\text{dim}}}$$

满足 Mercer 条件时，则用内积 $K(x,y)$ 代替最优分类面中的点积，即 $\Phi(x) \cdot \Phi(y) = K(x,y)$，相当于把原来的输入空间变换到新的特征空间，相应的基于支持向量机(support vector machine，SVM)的高层多模式融合函数为

$$f(x) = \text{sgn}\left(\sum_{i=1}^{n_{\text{dim}}} \alpha_i c_i K(s_i, s) + b\right) \tag{6}$$

其中，$\alpha_i > 0$ 为 Lagrange 系数，b 是分类的最佳阈值，可由任意一对支持向量取中值求得。sgn( )是符号函数。

3.3 融合训练算法

$X = \{x^{(1)}, \cdots, x^{(n)}\}$ 为训练样本集

$C = \{c_1, \cdots, c_n\}$ 为视频语义概念标签集

其中，$x^{(k)} = \{\{x_{1,1}^{(k)}, \cdots, x_{1,m_1}^{(k)}\}, \cdots, \{x_{n_{\text{dim}},1}^{(k)}, \cdots, x_{n_{\text{dim}},m_{n_{\text{dim}}}}^{(k)}\}\}$ 是样本 $x^{(k)}$ 中各个特征按 2.1 节进行的第 2 层模式划分的情况。其中 $x_{i,j}^{(k)}$ 表示第 i 类一级模式中的第 $j(j=1,2,\cdots,m_i)$ 类二级模式的特征，m_i 表示第 i 类一级模式中二级模式特征的数目，$k(k=1,2,\cdots,n)$ 为此样本在训练样本集中的序数。

两级融合算法描述：

(1) 低层模式融合(以第 i 类高层模式对应的数据为例)

① HHMM 模型训练　确定 HHMM 结构后，再将 X 中的特征向量 $x^{(k)} = \{\{x_{1,1}^{(k)}, \cdots, x_{1,m_1}^{(k)}\}, \cdots, \{x_{n_{\text{dim}},1}^{(k)}, \cdots, x_{n_{\text{dim}},m_{n_{\text{dim}}}}^{(k)}\}\}$(对应于第 i 组的多模低层特征)和类别标签集 $C = \{c_1, \cdots, c_n\}$ 作为 HHMM 输入，并按 3.1 节方法进行参数估计。

② 概率估计　用训练好的第 i 类 HHMM 模型 M_i^{HHMM} 来计算与样本 x_i 对应的各类别的似然率：$p_i = \{p_{i,1}, \cdots, p_{i,n_{\text{dim}}}\}$。

(2) 基于核的高层(级)融合

把低层融合所得的似然率 p_i 和类别标签 c 作为输入。先选择核函数，再通过训练 SVM，即可得到高层的融合函数 f。

两级融合训练算法：

Input: $X = \{x^{(1)}, \cdots, x^{(n)}\}$ /* 训练样本 */

$C = \{c_1, \cdots, c_n\}$ /* 视频语义标签 */

Output: f /* 多模式融合函数 */

Function Call:

$\{M_{1,1}^{\text{feature}}, \cdots, M_{1,m_1}^{\text{feature}}, \cdots, M_{n_{\text{dim}},1}^{\text{feature}}, \cdots, M_{n_{\text{dim}},m_{n_{\text{dim}}}}^{\text{feature}}\} \leftarrow \text{MP}(M_{\text{feature}})$ /* 视频数据训练样本中特征的两级模式划分，M_{feature} 表示特征模式 */

Train(x_i, C) /* 训练第 i 组多模特征融合的模型 */

$P(\lambda(M_i^{\text{HHMM}}, x_i))$ /* 将分类结果以似然率表示 */

TrainSVM(M, C) /* 训练高层的基于核的非线性融合模型，其中 M 表示确定的非线性融合模型 */

Procedure:

① $MP(M_{\text{feature}})$;

② for each $i = 1, \cdots, n_{\text{dim}}$

③ $M_i^{\text{HHMM}} \leftarrow$ Train(x_i, C); /* 低层融合 */

④ $p_i \leftarrow P(\lambda(M_i^{\text{HHMM}}, x_i))$;

⑤ $f \leftarrow$ TrainSVM(M, C); /* 高层融合 */

⑥ return f( );

4 实验分析

为验证本文算法的效果，选取了赛车等 8 类语义概念进行了不同融合方法语义分类效果的对比实验，实验中，所有实验样本选自 50h 的 MPEG 视频数据。样本按镜头为单位分割选取，每一类语义概念提取 200 个样本，其中 50% 用于训练，另外 50% 用于测试。多模式特征的提取按 MPEG-7 标准进行，即首先提取关键帧，并从关键帧中提取图像类的以下多模式特征量：可扩展颜色(scalable color)描述子、主颜色(dominant colors)描述子、均质纹理(homogenous texture)描述子、基于轮廓的形状(contour shape)描述子、边直方图(edge histogram)描述子；然后提取每个镜头的活动强度(intensity of activity)描述子；最后提取以下声音类多模特征：音频谱基(audio spectrum basis)描述子、音频谱投影(audio spectrum projection)描述子、音量(audio power)描述子、过零率(zero-crossing rate)；而文字类多模低层特征提取则按文献[9]中方法提取字幕。

HHMM 的层数为 3,2,1 等 3 种，不用的语义概念，按实际情况通过选取其中一种来确定其 HHMM 的层次结构。实验中核函数选取 3 阶多项式的形式。

表 1 是进行视频语义概念提取时，两级多模式

表 1　两级多模式融合算法与 PC、LC 直接融合方法对比实验

Tab.1　Comparison of experiment results for PC, LC and two level fusion method

编号	视频语义概念	直接单级融合方法语义分类准确率(%)		本文两级多模式融合算法				
		PC	LC	图像	声音	文字	n_{dim}	准确率(%)
1	赛车	78	85	4	2	1	3	93
2	飞机起飞/着陆	81	86	4	2	1	3	94
3	对话	90	93	0	2	0	1	96
4	户外	82	84	6	0	0	1	97
5	建筑	91	96	6	0	0	1	98
6	海浪	80	86	3	2	0	2	91
7	足球赛	84	85	3	2	1	3	92
8	爆炸场面	92	94	4	1	0	2	96

融合算法与常用的乘积融合(product combination, PC)和线性融合(linear combination, LC)两种直接融合方法的对比实验结果。其中,n_{dim} 表示视频语义概念提取时,两级多模式融合方法的高层融合的模式组数目,而图像、声音和文字则表示对应各模式组低层模式融合的模式数目。两种直接融合方法均采用单级融合,即将各低层模式特征作为一个独立的模式,采用典型的后期融合方式来进行融合。本文提出的方法在 $n_{dim}=2,3$ 时,实验是采用完整的两层多模式融合方法进行视频语义概念提取。而在 $n_{dim}=1$ 时,则不进行高层非线性融合,而是直接把低层多模式特征的融合结果作为最终值。对比的实验结果如图 4 所示。图中数据显示用本文中方法提取语义的准确率明显比 PC 和 LC 两种方法高。对比 PC 和 LC 两种方法的结果,从总体上看,LC 方法的融合效果比 PC 方法好,这是由于乘积融合对噪声有放大效应所致。由本实验的结果对比可得出,本文提出的两层多模式融合算法在提取视频语义概念时,不仅准确率较高,且可有效融合多模特征。

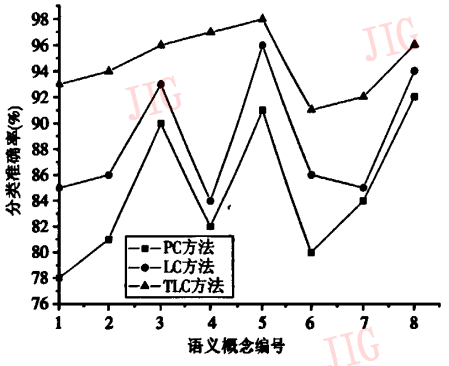

图 4　3 种融合方法结果

Fig.4　Comparison results of three methods

表 2 是在本文提出的两级融合框架下,高层融合采用不同方法的对比实验结果。3 种方法的低层融合均采用相同的 HHMM 结构,而高层融合则采用基于核的非线性融合(kernel-based combination, KC)、PC 与 LC 方法。n_{dim} 为高层融合的模式类别数目。从表 2 中的平均值看,KC 方法融合的语义概念提取准确率比乘积融合方法提高 5.2%,比线性融合方法提高 4.0%。图 5 是表 2 中 3 种高层融合方法提取的视频语义概念准确率的对比图。由图 5 可见,在本文提出的两层多模式融合框架下,采用基于核的非线性高层融合策略得到的效果较好。

表 2　两层融合框架下高层融合策略对比实验

Tab.2　Comparison of experiment results for different high-level combination methods

编号	语义概念	高层融合策略语义分类准确率(%)			
		n_{dim}	PC 方法	LC 方法	本文 KC 方法
1	赛车	3	88	90	93
2	飞机起飞/着陆	3	90	92	94
3	海浪	2	88	86	91
4	足球赛	4	82	84	92
5	爆炸场面	2	92	94	96
	平均	—	88.0	89.2	93.2

表 3 是本文提出两级融合算法的高层融合分别采用三阶多项式、径向基($\sigma^2=0.3$)、Sigmoid($b=2$, $c=1$)作为核函数训练后的视频语义分类的正确率

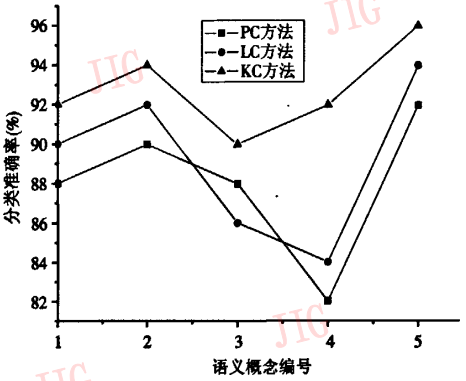

图 5 高层融合方法对比结果

Fig. 5 Comparison results Results of high level fusion method

对比实验结果。由表 3 数据可见,采用 3 种不同的核函数进行的高层融合,其融合分类结果与基于核的融合性能大致相同。可见该两级融合算法的高层融合对具体核函数类型的选择并不敏感。

表 3 高层融合不同核函数对比实验

Tab. 3 Comparison results for different kernel functions

语义概念	高层融合不同核函数语义分类准确率(%)		
	多项式	径向基	Sigmoid
赛车	93	93	92
飞机起飞/着陆	94	93	94
海浪	91	89	90
足球赛	92	93	92
爆炸场面	96	96	95

综合以上 3 个实验结果说明,文中提出的两级多模式融合方法充分利用了视频中多种模式的特征,不仅效果优于乘积融合及线性融合方法,并且对核函数类型不敏感。

5 结 论

在语义概念的提取过程中,本文提出了一种两级视频数据融合算法。该算法中,低层的融合采用 HHMM 模型,高层采用基于核的非线性融合。实验分析表明,该两级多模式融合算法不仅能有效融合视频多模特征,而且能准确提取视频高层语义信息,其性能较优。

参考文献(References)

1 Ganong W F. Review of Medical Physiology [M]. New York: McGraw-Hill publishing Company, 1999.

2 Xie L, Kennedy L, Chang S-F, *et al.* Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams[A]. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05) [C], Philadelphia, PA, USA, 2005, 2: 1053 ~ 1056.

3 Kuo J-H, Kuo J-B, Chen H-W, *et al.* A hierarchical and multi-model based algorithm for lead detection and news program narrative parsing[A]. In: Proceedings of the 19th International Conference on Advanced Information Networking and Applications(AINA'05)[C], Taipei, Taiwan, China, 2005: 511 ~ 514.

4 Wang F, Ma Y-F, Zhang H, *et al.* A generic framework for semantic sports video analysis using dynamic bayesian networks [A]. In: Proceedings of 11th International Conference on Multi Media Modeling (MMM 2005)[C], Melbourne, Australia., 2005: 115 ~ 122.

5 Wu Y, Chang E Y, Chang K C-C, *et al.* Optimal multimodal fusion for multimedia data analysis [A]. In: Proceedings of ACM International Conference on Multimedia (MM) [C], New York, USA, 2004: 572 ~ 579.

6 Fine S, Singer Y, Tishby N. The hierarchical hidden. Markov model: analysis and applications [J]. Machine Learning Engineering, 1998, 32(1): 41 ~ 62.

7 Müller K-R, Mika S, Rätsch G, *et al.* An introduction to kernel-based learning algorithms [J]. IEEE Transactions on Neural Networks, 2001, 12(2): 181 ~ 201.

8 Vapnik N. Hypostasis of Statistical Learning[M]. Zhang-Xue-gong Translate. Beijing: Tsinghua University Press, 2000(in Chinese). [Vapnik N 著,张学工译. 统计学习理论的本质[M]. 北京: 清华大学出版社, 2000.]

9 Shi Ying-chun. Study of Content Based Semantic Extraction Issertation [D]. Nanjing: Nanjing University of Science and Technology 2005.(in Chinese)[史迎春. 基于内容的视频检索语义提取若干问题研究[D]. 南京:南京理工大学, 2005.]