基于局部和全局特征视觉单词的人物行为识别

谢 飞 龚声蓉 刘纯平 季 怡
(苏州大学计算机科学与技术学院 苏州 215006)

摘 要 基于视觉单词的人物行为识别由于在特征中加入了中层语义信息,因此提高了识别的准确性。然而,视觉单词提取时由于前景和背景存在相互干扰,使得视觉单词的表达能力受到影响。提出一种结合局部和全局特征的视觉单词生成方法。该方法首先用显著图检测出前景人物区域,采用提出的动态阈值矩阵对人物区域用不同的阈值来分别检测时空兴趣点,并计算周围的3D-SIFT特征来描述局部信息。在此基础上,采用光流直方图特征描述行为的全局运动信息。通过谱聚类将局部和全局特征融合成视觉单词。实验证明,相对于流行的局部特征视觉单词生成方法,所提出的方法在简单背景的KTH数据集上的识别率比平均识别率提高了6.4%,在复杂背景的UCF数据集上的识别率比平均识别率提高了6.5%。

关键词 视觉单词,显著图,3D-SIFT,动态阈值矩阵,光流直方图
中图法分类号 TP391.4 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2015.11.060

Human Action Recognition by Visual Word Based on Local and Global Features

XIE Fei GONG Sheng-rong LIU Chun-ping JI Yi
(School of Computer Science & Technology, Soochow University, Suzhou 215006, China)

Abstract Different from the method based on low-level features, the human action recognition based on visual word adds mid-level semantic information to features and then improves the accuracy of recognition. For complex background or dynamic scenes, the efficiency of visual words might deteriorate. We proposed a new method which is a combination of local and global feature to generate visual words. Firstly, our approach uses saliency map to detect the rectangles around human. And then inside these rectangles, 3D-SIFT will be calculated around interest points detected from dynamic threshold matrix to describe local features. We also added HOOF to describe the global motion information. These visual words provide the important semantic information in the video such as brightness contrast, motion information, etc. The performance of this method in action recognition can be improved 6.4% on KTH dataset and 6.5% on UCF dataset compared with state-of-the-art methods. The experiment results also indicate that our visual dictionary has more advantages in both simple background and dynamic scene than others.

Keywords Visual words, Saliency map, 3D-SIFT, Dynamic threshold matrix, HOOF

1 引言

近年来,随着主题模型从在文本分类中的成功运用[1,2]逐步扩展到图像和视频处理中的应用[3-12],基于视觉单词的人物行为识别成为了一个研究热点。在人物行为识别领域,基于视觉单词的方法将整个图像和视频看成是由多个不分词序的视觉单词所组成,其中的视觉单词提取了视频中的低层特征并且加入了语义信息,因此相比于纯粹使用低层特征的方法,基于视觉单词的方法在识别精度上可以取得更好的结果。

Fei-fei Li等人[3]首先对图像进行规则分块,然后统计分块中的灰度直方图并将其作为视觉单词,从而对图像进行分类。这种方法较为简单,容易实现,但是对于天空和海洋等具有大量相同灰度信息的图片,基于灰度直方图的视觉单词难以区分。Jun Yang等人[4]使用DOG(Difference of Gaussian)检测子检测出图像中的关键点,之后计算关键点的PCA-SIFT(Principal Component Analysis-Scale-Invariant Feature Transform)特征作为图像的视觉单词,这种视觉单词具有较好的抗噪性和尺度不变性,但是使用DOG检测方法无法获得足够多的兴趣点,从而影响到了后续的图像分类。Laptev[5]首先将二维图像中的Harris角点检测方法扩展到了三维空间,将视频中检测出的三维Harris角点作为视频的

到稿日期:2014-07-21 返修日期:2014-08-29 本文受国家自然科学基金:基于二型模糊概率图模型的多摄像头目标跟踪研究(61170124),基于显著性和信任传递的动态场景主题发现(61272258),基于深度学习的时序3D深度图动作语义理解(61301299),江苏省产学研联合创新资金(前瞻性联合研究项目):复杂场景下异常行为分析及其应用(BY2014059-14)资助。

谢 飞(1986-),男,硕士,主要研究方向为图像处理与模式识别,E-mail:464518013@qq.com;龚声蓉(1966-),男,教授,博士生导师,主要研究方向为图像处理与模式识别,E-mail:shrgong@suda.edu.cn(通信作者);刘纯平(1971-),女,教授,硕士生导师,主要研究方向为图像处理与模式识别,E-mail:cpliu@suda.edu.cn;季 怡(1972-),女,副教授,主要研究方向为图像处理与模式识别,E-mail:jiyi@suda.edu.cn。

时空兴趣点，建立以兴趣点为中心的时空立方体，并构建光流直方图和梯度直方图作为特征来生成视觉单词。这种方法可以很好地提取出兴趣点周围的运动信息，但是容易受到光照的影响，尤其在运动相对平缓的视频中检测出的兴趣点过少，不利于后期的处理。为了解决兴趣点提取数量的问题，Dollar[6]使用 Gabor 滤波和高斯滤波来检测视频中的兴趣点，在提取出兴趣点之后，生成基于 Cuboids 的视觉单词，提高了单词的鲁棒性，但该视觉单词因使用简单的空间立方体而不能很好地反映出兴趣点周围的运动变化，且容易受到场景变化的影响。Scovanner[7]将二维 SIFT 特征扩展到三维，使用三维 SIFT 算子作为视频的视觉单词，这种视觉单词可以更好地表示行为的三维时空信息，有效地减少了噪声和光照等因素的影响，但三维 SIFT 特征对于运动信息的表示有一定的欠缺，在遇到两个动作近似的情况下，往往得不到较好的结果。Chaudhry 等人[8]使用光流直方图作为视频的全局特征来生成视觉单词，该视觉单词计算简单，不会因为图像中运动人物尺度的变化而受到影响，并且这种单词无需前背景分割。但由于使用全局特征，该方法得到的视觉单词舍弃了视频中的细节信息。Xia Lu 等人[9]使用了的三维肢体连接点的直方图作为视频全局特征，使用 Kmeans 将特征聚类后生成视觉单词，该单词包含了动作间的三维位置关系，且该方法运算速度较快，可以实时运行，对视角方向具有不变性，但是需要特殊的三维视频。Shao Ling[10]使用行为的历史运动信息生成的形状作为时间特征，使用梯度相关图作为空间特征，这种方法可以很好地识别连续的动作，但是如果行为中有停顿，轨迹信息就不能很好地进行描述。Ullah 等人[11]提出了一种将基于部件的人物运动轨迹作为视觉单词的方法，该方法可以很好地表示复杂的人物行为，但是在训练过程中需要大量的标注样本。Kovashka 等人[12]使用光流直方图和梯度方向直方图作为基本特征，在这些特征中再加入兴趣点周围的邻域信息来生成整个特征，然后聚类生成单词，由于这些单词还包含了兴趣点的邻域信息，因此提高了识别的准确性。上述视觉单词生成方法从整个图像和视频中提取出特征，没有对前景和背景区分对待，且大多数方法仅仅使用单一的特征，不能很好地表征人物的动作，尽管在简单背景下能取得较好的结果，但是当场景较复杂、噪声较多时，生成的单词往往会受到背景信息的干扰，影响之后的识别精度。

针对这些问题，本文提出使用显著图先得到人物大致区域，对区域内外采取不同的阈值来计算兴趣点，尽可能降低背景区域对兴趣点检测的影响，提高人物区域内提取出的兴趣点数量。进行低层特征选取时，结合局部 3D-SIFT(3D Scale-Invariant Feature Transform)特征和全局 HOOF(Histograms of Oriented Optical Flow)特征来描述人物的行为，得到的特征不仅具有 3D-SIFT 特征良好的旋转不变性和抗噪性，并且能够表征人物行为的全局运动信息。最后通过谱聚类[14]的方法将特征聚成视觉单词。简单场景 KTH 数据集和复杂场景 UCF 数据集上的对比实验表明，所提出的人物行为识别方法在各种场景中都能取得较好的结果。

2 提出的人物行为识别方法

图 1 给出了本文所提出的人物行为识别的框架结构，对于输入的视频，本文首先计算其对应的显著图来获取人物的大致区域，根据这一显著区域计算阈值矩阵并进行基于滤波的兴趣点检测，然后计算兴趣点周围的 3D-SIFT 特征来描述兴趣点周围的梯度信息，再计算整帧视频的 HOOF 特征并结合两个局部和全局特征，之后使用谱聚类将生成的特征聚成视觉单词，最后使用分类模型对视频进行行为的分类以达到识别的目的。

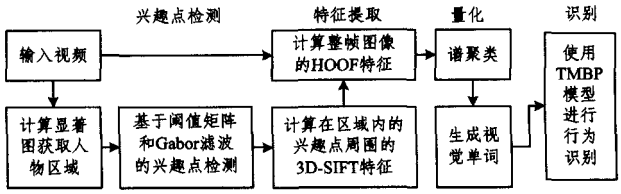


图 1 本文提出的人物行为识别框架

2.1 基于 GBVS 模型和动态阈值的兴趣点检测

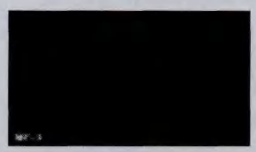
兴趣点的检测通常受到两个因素的影响。一是检测出来的兴趣点大多存在于背景区域，对前景内容的描述较少。由于显著图模型可以突出图像中的显著区域，降低背景的干扰，因此提出使用显著图的方式提取出显著的前景区域，然后对该区域提取兴趣点，以避免背景兴趣点对前景的干扰，提取出有利于后续行为识别的有效兴趣点。二是有效兴趣点数目对表征图像内容具有较大影响。数目太多会增加后续的计算复杂度，太少则又不能很好地表示场景内容，因此适当选择有效兴趣点成为另一个关键的问题。常见的基于滤波的兴趣点检测方法通常都使用单一的阈值来提取局部最大值，增大或减小该阈值会使得兴趣点的检测出现较多误差点，因此本文提出了动态阈值矩阵来解决这一问题。

2.1.1 基于 GBVS 模型的显著性检测

本文提出的视觉单词生成方法的第一步就是进行显著性检测。根据对显著性检测文献的分析，选择了性能比较好的 GBVS 模型。为了验证 GBVS 模型的效果，图 2 给出了几种显著性模型的显著图对比结果的一个示例。在显著度的计算过程中，复杂背景会影响得到的显著图，但是从实验对比图中可以看到，不同的方法对于背景的影响各不相同。图 2(a)是一组高低杠运动视频中的一帧图像，在整个视频中，图像背景区域中的人群始终存在鼓掌和站立的动作，相对于静止的背景，此类场景更为复杂。图 2(b)是基于剩余光谱的显著性模型的检测结果，虽然该模型可以勾勒出显著度的细节信息，但是从图中可以看出左下角有一片高亮区域，这是视频中的一个高亮时间标签，模型对亮度较为敏感，因此突出了该区域。图 2(c)是 PQFT 模型的运算结果，该模型该模型容易将近景的目标分割成多个显著区域，检测出的人物目标区域很小，并且背景区域和人物区域的显著度差距不大。图 2(d)是 Itti 模型的显著性检测结果，该模型在处理复杂场景时效果不好，背景干扰严重。图 2(e)是 GBVS 显著性计算的结果，可以看出，在原图中的人物位置所计算出的显著度值明显高于背景区域。图 2(f)是二值化图 2(e)后找出图像中的高亮部分而获取的人物大致区域。对比其他显著性方法，大量的实验分析证明，GBVS 模型更能突出场景中的前景目标，降低背景的影响，并且在预测人物位置时比其他方法更加准确可靠。因

此本文选择使用 GBVS 显著性模型对原始视频进行处理来获取前景目标区域。

(a)原始图像



(b)基于剩余光谱的显著图

(c)PQFT 显著图

(d)Itti 显著图

(e)GBVS 显著图

(f)人物大致区域

图 2　显著性模型对比结果

2.1.2　基于动态阈值的兴趣点检测

在视频处理中，基于兴趣点的方法通过计算兴趣点提取出视频的稀疏表示，无需对图像进行前背景分割和目标跟踪。常见的方法有基于角点的方法[5]、基于 LOG(Laplacian of Gaussian)的方法[18]和基于滤波的方法[6]。基于角点的方法将二维的角点扩展到三维空间，计算视频中的角点作为兴趣点。基于 LOG 的方法在检测兴趣点时使用高斯拉普拉斯作为响应函数。基于滤波的方法使用三维卷积窗口对整个视频进行卷积操作，然后寻找局部最大值作为兴趣点。前两种方法在检测兴趣点时，找到的兴趣点数量过于稀疏，不利于视频特征的提取，基于滤波的方法增加了兴趣点检测的数量，并且由于使用的是卷积操作，相对前两种方法更加简单，易于实现，时间复杂度更低。

因此本文使用基于 Gabor 滤波的兴趣点检测子[6]搜索图像中局部响应值最大的位置。该方法使用高斯滤波器在空间中对每一帧图像进行滤波，然后使用两个正交的一维 Gabor 滤波器在时间上进行滤波，之后定义响应函数：

$$R=(S*g*h_{ev})^2+(S*g*h_{od})^2 \tag{1}$$

其中，$g(x,y;\sigma)$是一个二维高斯平滑核，x 是横坐标，y 是纵坐标，S 是每一帧的输入图像，h_{ev} 和 h_{od} 是一对正交的一维 Gabor 滤波器，定义如下：

$$h_{ev}(t;\tau,\omega)=-\cos(2\pi t\omega)e^{-t^2/\tau^2} \tag{2}$$

$$h_{od}(t;\tau,\omega)=-\sin(2\pi t\omega)e^{-t^2/\tau^2} \tag{3}$$

σ 和 τ 是滤波器空间和时间上的两个尺度参数，$\omega=4/\tau$，t 是时间坐标。对于每个像素点，使用式(1)计算出它对应的响应值后，找出其中的局部最大值来作为整个视频的时空兴趣点。对于局部最大值的定义，文献[6]中的方法是使用统一的阈值来确定，但是这种方法在简单地调大阈值时，找到的兴趣点会非常少，而如果调小阈值，检测出的兴趣点虽然会增多，但是大部分的点都不能准确地定位到人物的动作上。因此本文根据 GBVS 显著图确定人物的大致区域，对区域内外使用不同的阈值，然后通过计算得到每一个像素点的阈值矩阵，再寻找局部最大值作为兴趣点。对于阈值矩阵的计算，首先定义空间上每个像素对应的阈值：

$$w_i=\begin{cases}\left(\dfrac{S+\delta}{S_{in}}\right)^{-1}\times\varepsilon_{in}, & \text{pixel in plane}\\ \left(\dfrac{S+\delta}{S_{out}}\right)^{-1}\times\varepsilon_{out}, & \text{pixel out plane}\end{cases} \tag{4}$$

其中，$S_{in}=\sum_{i=1}^{n}S_i$，S_i 是像素对应的显著度值，S_{in} 表示在区域内的所有像素的显著度值的总和。同样地，S_{out} 是区域外的所有像素的显著度值之和。δ 是一个微小的值，防止分母为 0。ε_{in} 和 ε_{out} 是两个权重因子，使区域内的权重总是比区域外的权重小。在时间上，计算连续的 $2\times\xi$ 的权重序列的平均值：

$$\overline{w_t}=\text{avg}\left(\sum_{i=t-\xi}^{t+\xi}w_i\right) \tag{5}$$

经过式(5)的计算后，就得到了一个三维的阈值矩阵。在后续的局部最大值的计算中，我们使用这个三维的阈值矩阵来代替单一的阈值。

2.2　3D-SIFT 和 HOOF 特征融合

SIFT 特征作为传统的特征描述子，具有尺度不变性、旋转不变性和光照不变性等特点。3D-SIFT 特征将 SIFT 特征从图像扩展到了视频中，可以较好地反映出兴趣点周围的梯度信息。而 HOOF 特征使用光流直方图描述了整帧图像的运动全局信息，可以很好地弥补 SIFT 特征缺少运动信息的缺点。因此本文使用 SIFT 特征和 HOOF 特征来描述图像兴趣点的局部和全局的信息。

2.2.1　3D-SIFT 特征

通过计算得到视频的时空兴趣点后，需要使用一种特征描述方法来表示兴趣点周围的时空信息。一种简单的方法是直接使用兴趣点周围的空间立方体结构作为特征[6]，这种方法比较简单，但是容易受到背景的影响。另一种常见方法是使用光流特征来描述视频[19]，但是由于光流计算速度较慢，并且对光照过于敏感，在光照有变化时，这种描述方法不利于后期的行为分析。

在实验中，本文使用候选点周围 2×2×2 的区域计算 SIFT 特征，对于梯度方向的规定，使用 20 个正三角形构建一个正二十面体。为了提高特征的表示，将每个面的正三角形再细分成 4 个正三角形从而构成一个正八十面体，将球心到每个正三角形的中心点的向量方向作为直方图的每个方向 bin，最终的特征长度为 2×2×2×20×4=640 维。

2.2.2 HOOF 特征

仅仅使用 SIFT 特征会存在一个问题，即物体的运动方向信息不能完全体现出来，并且 SIFT 特征是一种局部特征，无法获得整帧图像的信息，而光流特征可以很好地解决运动信息的问题。但是直接使用兴趣点周围的光流信息无法表示运动的整体信息，且容易受到噪声影响，因此使用光流直方图(HOOF)[8]特征来表示人物的运动信息。该特征是一个帧中所有光流的直方图，可以很好地避免光照等因素的影响，且可以作为该帧图像的全局特征，这样也可以弥补 SIFT 特征的局部性。

2.2.3 特征融合

本文将每个兴趣点的 3D-SIFT 特征和该兴趣点所对应的那帧图像生成的 HOOF 特征连接在一起，将每帧图像的局部特征和全局特征相结合。对于每一个兴趣点，由前文所述可以计算得到它的一个 640 维的 3D-SIFT 特征($x_1, x_2, \cdots, x_{640}$)，假设此兴趣点所在帧的全局 HOOF 特征为($y_1, y_2, \cdots, y_t$)，其中 t 为直方图的 bin 值。通过将两个特征拼接在一起得到一个新的特征($x_1, x_2, \cdots, x_{640}, y_1, y_2, \cdots, y_t$)。在后续的聚类生成视觉单词过程中，较大的 t 值会使得聚类中心更加偏向 HOOF 特征。因此，本文没有直接在特征融合时加入权重系数，而是通过调整直方图的 bin 值来调节全局特征在整个特征中所占的比重。

对于差别较大的行为，只使用 3D-SIFT 特征就可以很好地识别；而对于较为接近的行为，则需要在此基础上增加 HOOF 特征，因此在实验过程中，本文融合的特征以 3D-SIFT 特征为主，HOOF 特征为辅。在进行多次实验比较后发现，t 的值在 120～200 时可以得到较高的识别准确率。本文取 t=150 来进行实验对比。

2.3 谱聚类的视觉单词生成

谱聚类[14]是近年来比较流行的一种模型聚类方法，它利用了特征向量来求得问题近似解。谱聚类的本质是将聚类问题转化为图的最优划分问题，是一种点对聚类算法。它与普通的聚类方法的不同在于：普通的聚类算法直接对输入数据进行运算，而谱聚类首先对数据进行处理，计算出数据间的相似度矩阵，然后通过相似度矩阵得到拉普拉斯矩阵，并求解它的某些特征向量，然后对这些特征向量再进行聚类。由于谱聚类使用了基于图的拉普拉斯矩阵，因此只需要数据之间的相似度矩阵即可，而不要求数据必须是 N 维欧氏空间中的向量。相比于 Kmeans 等传统聚类算法，谱聚类易于实现并且能取得更好的精度和速度。文献[20]在 TDT2 数据集上对各种聚类算法做出了比较，并且证明谱聚类在文本聚类中的准确性优于大部分聚类算法。在视频处理中，视觉单词对应文本处理中的单词，视频对应于文章，在聚类操作时两者差异不大，因此本文将谱聚类作为视觉单词的聚类方法。

谱聚类的大体流程如下：给定一个数据点集 $X_1, \cdots, X_n$，定义相似度矩阵 S，其中 S_{ij} 表示 X_i 和 X_j 之间的相似性。非归一化的拉普拉斯矩阵定义为 $L=D-S$，其中 D 是一个对角矩阵，$D_{ii}=\sum_{j=1}^{n} S_{ij}$。

Step1 计算相似性矩阵 $S \in R^{n \times n}$；

Step2 计算非归一化拉普拉斯矩阵 L；

Step3 计算 L 矩阵的前 k 个特征向量 $u_1, \cdots, u_k$；

Step4 构造一个矩阵 $U \in R^{n \times k}$，其中每一列是一个向量 $u_1, \cdots, u_k$；

Step5 使用 Kmeans 聚类算法对矩阵 U 进行聚类计算，得到聚类中心。

在实验中，本文使用欧氏距离作为衡量相似性的标准来构建相似性矩阵，通过谱聚类对生成的特征进行聚类后，使用聚类中心作为视觉单词。在数据训练过程中，将聚类中心作为单词，使用分类模型进行分类建模；在数据测试过程中，计算测试视频中的所有融合特征，之后将每一个融合特征与训练过程中的聚类中心进行比较，选择距离最近的聚类中心作为该特征的视觉单词表示，最后再利用训练好的分类模型对测试视频进行分类。

3 实验结果与分析

针对所提出的视觉单词，本文使用简单场景的 KTH 数据集和复杂场景的 UCF 数据集来进行测试。从图 3 和图 4 中可以看出，KTH 数据集中的场景非常单一并且背景没有较大的改变，而 UCF 数据集存在大量的复杂和动态背景。

3.1 动作数据集

整个实验对两个行为数据集进行了分类测试，分别是从 KTH 数据集以及 UCF 数据集中取出的 6 种具有代表性的动作集。

KTH 数据集包括 boxing、handclapping、handwaving、jogging、running 和 walking 几个动作，由 25 个人物在 4 种场景下完成，每个动作有 100 个视频，共计 600 个视频。

从 UCF 数据集中选出的 6 种动作分别为：diving、horse riding、lifting、swing bench、swing sideangle 和 tennis。其中跳水和骑马为动态背景，举重和鞍马为简单背景，网球和高低杠为复杂背景，并且跳水、鞍马、高低杠和网球等几个动作包含很多人物旋转的特性。

3.2 实验结果分析

在整个实验中，本文首先使用 GBVS 对整个视频计算显著性图像，之后使用区域生长的方式确定人物的大致位置，在寻找时空兴趣点时，沿用文献[6]中通过参数取值与分类误差率的实验得到的最优参数取值，以 $\sigma=2, \tau=4$ 为尺度来寻找时空兴趣点。在生成阈值矩阵时，本文通过人物区域内外兴趣点数量的平均差值来评估参数的取值，最终确定为 $\delta=e^{-6}, \varepsilon_{in}=10^{-6}, \varepsilon_{out}=10^{-3}$。然后通过对显著图计算 3D-SIFT 和 HOOF 特征来建立时空单词。最后对这些单词使用谱聚类的方法进行聚类构建码本。

3.2.1 使用显著图和阈值矩阵与传统兴趣点检测的比较

图3和图4分别给出了使用统一阈值与使用显著图和阈值矩阵在KTH和UCF数据集上检测兴趣点的结果。左侧的图像为使用单一阈值的结果，从图中可以看出：检测到的兴趣点并非集中在人物周围，尤其是在UCF数据集上，有很多兴趣点被复杂背景所干扰。而右侧的图像是使用阈值矩阵的实验结果，从图中可以看出：对区域内外使用不同的阈值可以减少背景区域检测出的误匹配点，且仅在人物周围增加兴趣点的数量而非全局增长。

(a)boxing 使用单一阈值

(b)boxing 使用阈值矩阵

(c)jogging 使用单一阈值

(d)jogging 使用阈值矩阵

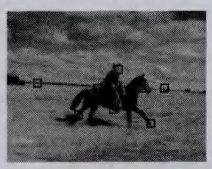
图3 在KTH数据集上使用单一阈值与使用阈值矩阵检测兴趣点的结果

(a)diving 使用单一阈值

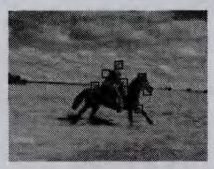
(b)diving 使用阈值矩阵

(c)swing bench 使用单一阈值

(d)swing bench 使用阈值矩阵



(e)horse riding 使用单一阈值



(f)horse riding 使用阈值矩阵

(g)swing sideangle 使用单一阈

(h)swing sideangle 使用阈值矩阵

图4 在UCF数据集上使用单一阈值与使用阈值矩阵检测兴趣点的结果

3.2.2 使用不同特征构建视觉单词的比较

表1是使用几种不同特征构建视觉单词后，统一使用SVM分类器进行行为识别的平均准确率的对比结果。

表1 几种视觉单词在人物行为识别上准确率的对比结果

Method	Accuracy(%)	
	KTH	UCF
Laptev[5]	92.00%	75.90%
Dollar[6]	84.70%	71.50%
Scovanner[7]	89.40%	77.50%
Chaudhry[8]	91.60%	74.60%
Ke[21]	80.90%	—
Niebles[22]	82.60%	—
Willems[23]	84.20%	—
Guha[24]	90.50%	79.60%
Baysal[25]	90.70%	74.80%
本文的方法	93.70%	82.20%

Laptev检测出视频中的局部三位Harris角点，然后提取周围的梯度信息生成单词来进行识别，这种方法在简单场景下的分类结果较好，但是在复杂场景下检测的兴趣点数量不够，无法很好地为后续的分类进行建模。Dollar检测出时空兴趣点后，提取兴趣点周围的立方体作为视觉单词，该方法实现简单，但是在动态场景中，受背景影响较为明显，难以取得较好的结果。Scovanner提取兴趣点周围的3D-SIFT特征来描述行为，这种特征维度大，无法较好地描述行为的运动信息，在区分jogging和running两类较为接近的行为时有困难。Chaudhry的方法中特征提取的是整帧图像的全局光流运动特征，对于局部变化不能较好地捕捉。Ke基于空间形状体来描述整个行为，这种方法需要前期对图像进行分割和检测，因此在场景复杂、分割精度下降的情况下，最终的检测结果会比较差。Niebles使用了线性分类器来改进兴趣点的提取方法以增加兴趣点的数量，但提取的特征仍然与Dollar的类似，在后期分类时，他们采用了主题模型的方法，但是由于视觉单词描述没有较大改进，分类结果的提高不明显。Willems的方法也是通过改变空间和时间尺度两个因素来提高兴趣点的稠密程度，但是在增加数量的情况下，精度上并没有太大提高，因此未能取得很好的结果。Guha将每个视频进行分段处理，对每一段的第一帧提取二维的兴趣点检测，之后根据这些兴趣点来提取每一段的空间立方体结构，之后提取出立方体结构中的时空特征，这种方法可以很好地描述时间尺度的信息，但是很难保证每一段第一帧视频中的动作具有代表性。Baysal首先提取出每一帧图像的线条，过滤掉背景的干扰，得到人物的轮廓，之后简化线条的表示，并对线条进行配对来描述该帧人物的姿态，最后对每一帧提取出全局的线条流直方图信息来表示行为，这种方法同样容易受到背景的干扰，在复杂场景下很难清晰地描述出人物的线条特征来。

3.2.3 不同单词数对实验结果的影响分析

图5示出了不同的视觉单词数量对行为识别召回率的影响。其中实线部分是在UCF数据集上的召回率，虚线部分是在KTH数据集上的召回率。整体的召回率以简单场景的KTH数据集为优。在实验中，以500个单词数量为单位，逐渐增加，可以看出，当单词数量过多或过少时，结果均不是最优。单词数量过少，就会忽略特征间的细节，不能够充分描述人物的行为；单词数量过多，单词与单词之间就会存在大量的冗余信息，也会影响最后的识别结果。在对KTH数据集进行聚类时，一共得到45万个特征向量，使用谱聚类将这些特征聚成2000个视觉单词时，效果最好。在对UCF数据进行

聚类时，一共得到23万个特征向量，视觉单词数量在1500个时可以获得最佳结果。

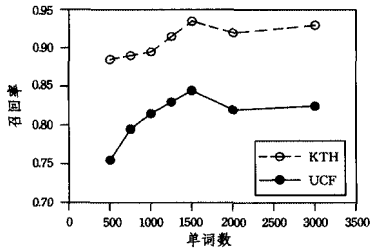


图5 在KTH和UCF上单词数对召回率的影响

结束语 本文提出了一种基于局部和全局特征的时空单词的人物行为识别方法。不同于常见的基于词袋模型的人物行为识别方法，本文提出了一种新的词袋生成方法，该词袋中的视觉单词考虑到全局和局部两方面，在局部使用3D-SIFT特征来描述行为细节的梯度信息，而在全局则使用HOOF特征来描述行为的全局运动信息，这种视觉单词具有较好的光照不变性、抗噪性和旋转不变性，并且能够充分地表征人物的行为。在提取兴趣点的过程中，本文通过利用显著性模型来得到人物的大致区域，之后在区域内外使用不同的阈值矩阵来减弱背景区域对兴趣点检测的影响，提高了兴趣点检测的精度，进而提高了后续的视觉单词的表征能力。在实验对比中发现，相对于几种基于词袋的人物行为识别方法，本文提出的识别模型在简单和复杂动态场景中均能得到较高的识别精度。

参 考 文 献

[1] Hofmann T. Probabilistic latent semantic indexing [C]// ACM SIGIR Conference on Research and Development in Information Retrieval. 1999:50-57

[2] Blei D M. Probabilistic models of text and images[D]. California: University of California, 2004

[3] Li Fei-fei, Perona P. A bayesian hierarchical model for learning natural scene categories [C]// 2005 IEEE Computer Vision and Pattern Recognition. 2005:524-531

[4] Yang J, Jiang Y G, Hauptmann A G, et al. Evaluating bag-of-visual-words representations in scene classification [C]// International Workshop on Multimedia Information Retrieval. 2007: 197-206

[5] Laptev I. On space-time interest points [J]. International Journal of Computer Vision, 2005, 64(2/3): 107-123

[6] Dollár P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features [C]// 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. 2005:65-72

[7] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition [C]// Proceedings of the 15th International Conference on Multimedia. 2007:357-360

[8] Chaudhry R, Ravichandran A, Hager G, et al. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2009: 1932-1939

[9] Xia L, Chen C C, Aggarwal J K. View invariant human action recognition using histograms of 3d joints [C]// 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. 2012:20-27

[10] Shao L, Ji L, Liu Y, et al. Human action segmentation and recognition via motion and shape analysis [J]. Pattern Recognition Letters, 2012, 33(4): 438-445

[11] Ullah M M, Laptev I. Actlets: A novel local representation for human action recognition in video [C]// 2012 19th IEEE International Conference on Image Processing. 2012:777-780

[12] Kovashka A, Grauman K. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition [C]// 2013 IEEE Conference on Computer Vision and Pattern Recognition. 2010:2046-2053

[13] Harel, Jonathan, Koch C, et al. Graph-based visual saliency [C]// Proceedings of the 20th Annual Conference on in Neural Information Processing Systems. 2006:1523-1527

[14] Von Luxburg U. A tutorial on spectral clustering [J]. Statistics and Computing, 2007, 17(4): 395-416

[15] Hou X, Zhang L. Saliency detection: a spectral residual approach [C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. 2007:1-8

[16] Guo C, Zhang L. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression [J]. IEEE Transactions on Image Processing, 2010, 19(1): 185-198

[17] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(11): 1254-1259

[18] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110

[19] Efros A A, Berg A C, Mori G, et al. Recognizing action at a distance[C]// 9th IEEE International Conference on Computer Vision. 2003:726-733

[20] Cai D, He X, Han J. Document clustering using locality preserving indexing [J]. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(12): 1624-1637

[21] Ke Y, Sukthankar R, Hebert M. Spatio-temporal shape and flow correlation for action recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2007:1-8

[22] Niebles J C, Wang H, Li Fei-fei. Unsupervised learning of human action categories using spatial-temporal words [J]. International Journal of Computer Vision, 2008, 79(3): 299-318

[23] Willems G, Tuytelaars T, Van Gool L. An efficient dense and scale-invariant spatio-temporal interest point detector [M]// Computer Vision-ECCV 2008. Springer Berlin Heidelberg, 2008:650-663

[24] Guha T, Ward R K. Learning sparse representations for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence 2012, 34(8): 1576-1588

[25] Baysal S, Duygulu P. A line based pose representation for human action recognition [J]. Signal Processing: Image Communication, 2013, 28(5): 458-471