基于稀疏编码和局部时空特征的人体动作识别

朱岩　赵旭　刘允才

上海交通大学图像处理和模式识别研究所，上海， 200240

摘　要　利用从视频中提取局部时空特征，许多最近提出的动作识别算法取得了较好的性能。这些方法通常使用Bag of Words（BoW）模型作为视频的表达方法。然而，BoW模型粗略地将每个特征向量量化为与之距离最近的一个视觉字，因此不可避免地带来较大的量化误差，从而妨碍了识别率的进一步提升。在本文中，为了得到更为精确，更具判别能力的表达形式，我们提出了一种新颖的动作识别方法：在稀疏编码的框架下，对局部三维时空梯度特征进行编码。通过这一过程，将每个局部时空特征变换为预先训练好的字典中的若干个“基”的线性组合。此外，根据迁移学习，我们探索了字典的建立方法。我们从各种体育比赛和电影中收集了大量视频片段，并通过在线学习的方法，从中训练出一组通用的基元作为字典。我们在KTH数据集和UCF体育数据集上测试该算法。实验结果表明，我们的方法在KTH数据集上超过了当前已发表的最先进的方法；在UCF体育数据上也取得了与先进水平相当的分类结果。

关键词　动作识别　稀疏编码　迁移学习　时空特征

中图法分类号：　**文献标识码：**　**文章编号：**

Sparse Coding on Local Spatial-Temporal Volumes for Human Action Recognition

ZHU YAN, ZHAO XU and LIU YUNCAI

Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, 200240

Abstract By extracting local spatial-temporal features from videos, many recently proposed approaches for action recognition achieve promising performance. Bag of words (BoW) model is commonly used in the approaches to obtain the video level representations. However, BoW model roughly assigns each feature vector to its closest visual word, therefore inevitably causing nontrivial quantization errors and impairing further improvements on classification rates. To obtain a more accurate and discriminative representation, in this paper, we propose a novel approach for action recognition by encoding local 3D spatial-temporal gradient features within the sparse coding framework. In so doing, each local spatial-temporal feature is transformed to a linear combination of a few “atoms” in a trained dictionary. Besides, we also explore the construction of the dictionary under the guidance of transfer learning. We collect a large set of diverse video clips from various sports games and movies, from which a set of universal basic atoms composed of the dictionary, are learned by an online learning strategy. We test our approach on KTH dataset and UCF sports dataset. Experimental results demonstrate that our approach outperforms the state-of-art techniques on KTH dataset and achieves the comparable performance on UCF sports dataset.

Keywords　action recognition, sparse coding, transfer learning, dictionary training

0 引 言

人体动作识别对于包括人机交互、智能监控和自动视频标注在内的多种应用具有重要价值。然而由于在真实场景中存在着巨大的类内差异，例如不同的人体姿态、遮挡和视角差异等，精确的动作分类依然是一个极具挑战性的问题。

近年，一些基于局部时空描述子[1,2,3]的动作识别方法取得了较好的结果。这些方法一般先从视频中检测或者密集采集出一组时空兴趣点，然后再以这些兴趣点为中心建立小立方体，并描述其局部统计特性。为了从局部特征中得到全局表达形式，这些方法多数采用了流行的 Bag of Words(BoW)模型，用来与各种局部时空描述子相配合。在 BoW 模型中，输入视频被看作大量“时空单词”的无序集合。通过在特征空间中测量距离尺度，每个“时空单词”被量化为已训练字典中的某个“单词”。最终，输入视频表示为字典中所有单词出现频率的直方图统计。然而，BoW 模型的量化策略中存在着一个缺陷：仅仅将每个特征量化为与之最接近的“单词”，这可能会导致相对较大的量化误差，因此得到的原始特征的近似表达过于粗糙，在之后的分类过程中不能提供足够的判别信息。

为解决这一问题，我们提出了一种新颖的基于稀疏编码框架的动作识别方法。系统流程如图 1 所示。首先，我们从输入视频中密集提取一组不同尺度下的局部时空特征。在试验中，我们采用了 HOG3D[2]描述子。然后，我们根据事先训练好的字典，使用稀疏编码的方法，将每个局部特征子转换为与之对应的稀疏码。在视频的整个稀疏码集合中进行最大值合并，从而得到了视频的最终表达。在分类过程中，我们使用了多类线性支持向量机。

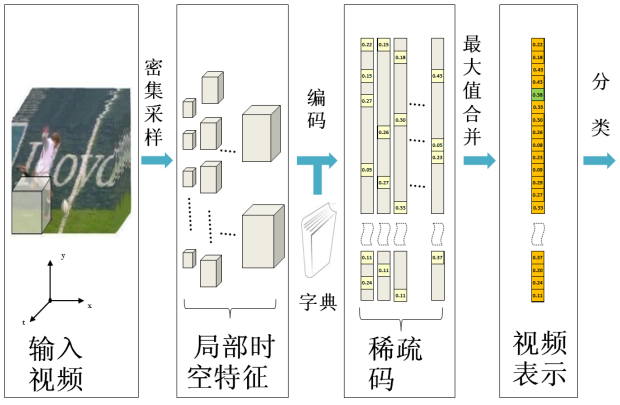

图1 本文动作识别系统的结构图。首先输入视频通过密集采样，转化为一组局部时空特征；然后根据预先训练好的字典，将局部时空特征编码为稀疏码集合；最终，在整个稀疏编码集合上进行最大值合并，得到最终的视频表达。

Fig1 The framework of our action recognition system. First the input video is transformed to a group of local spatial temporal features through dense sampling. Then the local features are encoded into sparse codes using the pre-trained dictionary. Finally, max pooling operation is applied over the whole sparse code set to obtain the final video representation.

本文强调的另外一个问题是，如何有效地对给定特征空间中局部时空特征的分布建模。在稀疏编码框架下，如何建立一个好的字典对于这一问题非常关键。通常，训练一个字典需要大量标签已知的训练视频集合。然而，为了得到一个更为普适的字典，需要大量标记过的视频数据。可是在大型视频数据库中进行视频标注是异常困难的。此外，使用同源的训练和测试数据会弱化字典的推广能力。针对此问题，近期提出的一些图像分类的工作[7,8]引入了迁移学习的概念。在这些工作的启发下，我们在试验中使用大量未标记的视频片段来建立字典。这些视频片段采样自电影和体育比赛中。尽管这些未标记的视频可能在语义层面上与待分类的动作不甚相关，但是字典可以从中学习，得到特征空间中分布的更为一般化的隐含结构。

1 相关工作

当前许多动作识别的方法通过检测一组特征点来得到视频的表示[9,10,11]。为描述这些特征点

的属性，提出了各种局部时空描述子。尽管特征点具有计算简便，表达紧凑的优点，这种方法遗弃了大量可能有用的原始信息，因此弱化了表达的判决能力。近期，两份评估确认了密集采样在动作识别中可以得到更好的结果[12,13]。在本文中，我们也采用了密集采样的方法，但是与上述不同之处在于，我们采用了稀疏编码，得到了一种新颖的视频中层表示形式。

稀疏表达在近年得到了广泛关注，并在许多领域取得了显著的成绩，包括语音识别[14]，图像恢复[15]，分割[16]，和物体识别[17,18]。这些工作的成功表明：稀疏表达可以灵活地拟合多种自然信号，并带有比较理想的特性。此外，大脑视觉皮层的研究也在生理学上找到了稀疏编码的依据[19]。对于人体动作分类的任务，所感兴趣的信息在本质上是在时空域上稀疏分布的。基于上述观点，我们在本文中将稀疏表达引入了动作识别这个领域。

为提高识别率，我们将迁移学习包含进稀疏编码的框架中，从未标记数据中获得知识。Raina等人[8]提出了自学习，利用未标记的语义无关数据建立稀疏编码字典，应用于图片分类的任务中。Yang 等人[17]将稀疏编码和尺度金字塔模型整合，应用于物品分类中。Liu 等[20]提出了一种拓扑子空间模型，利用归纳迁移学习，并应用于图像分类和检索中。受上述方法的启发，我们在本文中探索了稀疏编码和迁移学习在视频领域中的应用。实验结果证实，从未标记数据中得到的可迁移知识可显著提高动作识别率。

2 方 法

2.1 局部描述子和采样策略

为得到足够的具有较强判别能力的信息，我们在输入视频的时空域内密集采样，得到一个局部特征集合。每个描述子根据以每个时空点为中心的三维立方体内计算而得。为增强尺度和速度的不变性，我们利用了多尺度三维结构。为获得局部运动和外形特征，我们在试验中使用了HOG3D 描述子[2]。

2.2 稀疏编码

在我们的动作识别系统中，稀疏编码的作用是对人体动作得到更具判别能力的中层表达。假设我们已经得到一组局部时空特征向量 $X=[X_1,\cdots,X_n]$，用来表示一个视频片段，其中每个特征都是一个 d 维列向量。此外，假设我们已经得到一个训练好的字典矩阵 $D=[d_1,\cdots,d_S]$。稀疏编码的作用是，将每个特征向量稀疏的编码为字典 D 中的一些“基”的线性组合的形式。这一编码的实现是通过优化下述公式：

$$\hat{Z}=\arg\min\frac{1}{2}\left\|X-DZ\right\|_{l_2}^2+\lambda\left\|Z\right\|_{l_1} \qquad (1)$$

其中 λ 是正则化系数，它决定了表达形式的稀疏性。字典 D 是一个预先训练好的过定的“基”的集合，由 S 个“基”组成。为了避免数值不稳定，D 中的每一列都满足 $\left\|d_k\right\|_{l_2}\le 1$ 的约束。当字典 D 是固定的，则对 Z 的优化是凸的，因而可以看做是一个 1 范数正则化线性最小平方问题。这一优化同时保证了最小重建误差和系数矩阵的稀疏性。针对此问题，我们使用 LARS-lasso 的方法[15]得到最优解。经过优化之后，我们得到了一组稀疏码 $\hat{Z}=\left[\hat{z}_1,...,\hat{z}_N\right]$，其中每个列向量 $\hat{z}_i$ 只有若干个非零元素。

为得到整个视频的全局统计特性，我们使用了以下的最大值合并方程：

$$\beta=\xi_{\max}\left(\hat{Z}\right) \qquad (2)$$

来对整个稀疏码集合 Z 进行合并，其中 $\xi_{\max}$ 函数返回一个 S 维向量 β，它的第 k 个元素定义为

$$\beta_k=\max\left\{\left|\hat{Z}_{k1}\right|,\left|\hat{Z}_{k2}\right|,...,\left|\hat{Z}_{kN}\right|\right\}。 \qquad (3)$$

通过最大值合并，得到的向量 β 即为最终的视频级特征。最大值合并操作已经被成功应用于若干个图片分类的系统中，用来提升空间变换不变性[17, 18, 21]，并且已经被证明与视觉大脑皮层的细胞特性一致[21]。在这一观点的启发下，我们采取了相似的方法，用来提高空间和时间变换的不变性。在整个稀疏码集合中，我们只保存每个“基”的最强烈反应，而不考虑它的时空位置。试验结

果表明，最大值合并可以有效地带来紧凑和较高判别性的视频表示。值得注意的是，之前所采取的密集采样策略为这一步的操作提供了充足的低级局部特征，从而保证了最大值合并操作在统计上的可靠性。

2.3 基于迁移学习的字典建立方法

在稀疏编码的框架下，建立特定分类任务的字典，本质上是去学习一组过定的“基”，用来模拟特定数据在特征空间内分布的基本模式。给定一组大量的局部描述特征 $Y=[y_1,...,y_M]$，字典学习的过程实际上是对字典 D 和系数集合 $Z=[z_1,...,z_M]$ 联合优化的过程，如下：

$$\min_{Z\in\ ^{S\times M},D\in C}\frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}\|y_i-Dz_i\|_{l_2}^2+\lambda\|z_i\|_{l_1} \quad (4)$$

其中 C 是一个凸集：

$$C\quad\left\{D\in\ ^{d\times S},s.t.\|d_k\|_{l_2}\le 1,\forall k\in\{1,...,S\}\right\}。$$

在字典训练阶段，字典 D 和系数集 Z 同时变化，所以此优化问题是非凸的。如何解这一优化问题，尤其是面临海量的训练数据，依然是一个尚未定论的问题。最近，Mairal 等人[15]提出了一种新的在线字典训练的算法，这一算法收敛速度较快，更适合于动作识别的大规模训练数据。由于上述特性，我们采取了这种方法来训练我们的字典。

为了更好的发现特征空间内的隐藏结构，我们使用了大量未经标注的视频片段。这些片段采集自电影和体育比赛中，作为“学习材料”。近年来，从未标注数据中迁移学习，使用到监督分类中，已经激发出许多成功的应用[8,20]。尽管这些未标注视频片段与测试数据并不一定来自相同的分类中，但是在人体运动中，都包含通用的基本模式。这一可迁移的知识对我们的有监督动作分类是有益的。

在实验中，我们用不同的方式建立了两个字典。第一个字典使用目标分类视频片段训练，而第二个使用未标记视频片段训练。实验结果表明，第二个字典得到了更高的识别率。我们对这一差别做出如下解释：未标记的数据包含有更具多样性的人体动作的模式，这有利于字典挖掘到人体动作的更本质的属性。与之对比，由于训练数据的相对单一化，仅通过训练数据建立的字典难以得到较通用的“基”。值得注意的是，通过未标记数据训练的字典可以应用于其他相关的动作分类任务中。知识迁移使得我们的字典具有更好的推广性和可复用性。

3 实 验

我们在两个通用的人体动作数据库上测试了我们的算法：KTH 动作数据库和 UCF 体育数据库。

3.1 参数设定

字典训练参数：在字典训练过程中，我们经验地将字典大小设为 4000。在字典学习过程中，我们从 500 个视频片段中抽取 400000 个 HOG3D 特征。在选择视频片段的过程中，我们没有对视频内容加入任何语义上的约束，唯一的要求是：视频中必须包含运动的人体。根据文献[15]的建议，l_1 正则化参数 λ 设定为 $\frac{1.2}{\sqrt{m}}$，其中 m =960，代表原始信号的维数。

3.2 KTH 数据库的实验结果

KTH 数据库是一个评估各种动作识别算法的标准数据库。它包含 6 种人体动作，包括行走、慢跑、跑步、拳击、挥手和拍手。每个动作各被 25 个表演者在 4 种不同的环境下重复若干次。KTH 数据库共包含 599 个视频片段。我们采用通常的实验方法，即随机将所有视频分成训练集（16 个表演者）和测试集（剩余 9 个表演者）。在分类时，我们使用多类支持向量机，采用了一对多的策略。相同的实验重复 100 次，并记录所有类的平均准确率。

表 1　KTH 数据库的实验结果与已发表文献结果的对比

Tab.1 Comparisons to previous published result on KTH dataset

方法	平均准确率	实验设置

Niebles[4]	81.50%	留一法
Jhuang[23]	91.70%	分割法
Fathi[24]	90.50%	分割法
Laptev[1]	91.80%	分割法
Bregonizio[25]	93.17%	留一法
Kovashka[6]	94.53%	分割法
Our Method	**94.92%**	**分割法**

表 1 表明，我们的方法取得了 94.92%的准确率，这一结果超出了所有目前公开的分类结果。图 2 展示的矩阵表明，除慢跑和跑步以外，所有类别都取得了令人满意的准确率。这一误差是可以理解的，因为这两种动作即使对人眼来说也容易混淆。

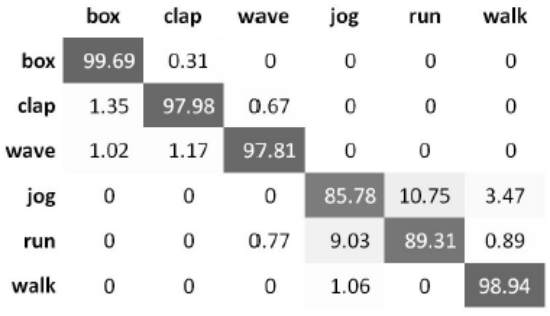

图2　在KTH数据库上实验结果的混淆矩阵，行标签为真实标签，列标签为预测标签。所有报告的结果为100次重复实验的平均。

Fig2　Confusion matrix for KTH dataset, the horizontal rows are the real labels while the vertical columns are predicted ones. All the reported results are the averages of 100 rounds.

3.3 UCF 体育数据库的实验结果

UCF 体育数据库包括 150 个视频片段，分属 10 类动作，包括跳水、挥高尔夫球棒、踢、举重、骑马、跑步、滑板、旋转（绕单杠）和旋转（鞍马）。所有上述视频均采集自现实体育节目中。与[13,22]类似，我们将所有原始视频左右翻转。为了公平的与先前公开的结果相对比，我们也采用“留一法”，即测试每个原始视频，并同时把所有其他视频作为训练集。翻转过的视频不作为训练集。我们以各类平均准确率作为结果。实验结果表明，我们的方法取得了与先进水平相当的准确率，如表 2 所示。图 3 给出了 UCF 数据库的混淆矩阵。

表 2　UCF 体育数据库的实验结果与已发表文献结果的对比

Tab.2 Comparisons to previous published result on UCF Sports dataset

方法	平均准确率	实验设置
Rodriguez[22]	69.30%	留一法
Yeffet[26]	79.30%	留一法
Wang[13]	85.60%	留一法
Kovashka[6]	87.27%	留一法
Our Method	**84.33%**	**留一法**

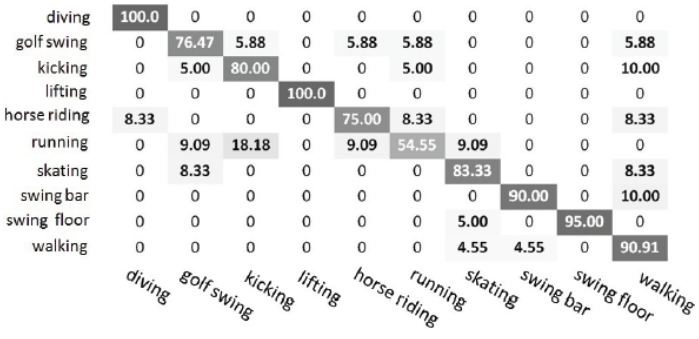

图3　在UCF体育数据库上实验结果的混淆矩阵

Fig3　Confusion matrix for UCF Sports dataset

4 结 论

在本文中，我们提出了一种新颖的动作识别方法，使用稀疏编码和局部时空描述子。为得到高层视频级的表达，我们使用了稀疏编码和最大值合并，用来取代传统的 BoW 直方图统计表示。实验结果表明，稀疏编码可以提供更为准确的表示，并具有理想的稀疏特性。这一优点有助于加强判别能力，从而提升识别准确率。在未来的工作中，我们将研究字典的内在结构和不同“基”的相互关系。此外，监督字典学习算法也将是我们未来的研究方向。

参考文献

[1] Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B. Learning realistic human actions from movies [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2008:1 – 8

[2] Klaser, A., Marszalek, M., Schmid,C. A spatio-temporal descriptor based on 3D-gradients [C]// British Machine Vision Conference. 2008

[3] 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, ACM. 2007:357-360

[4] Niebles, J.,Wang, H., Fei-Fei, L. Unsupervised learning of human action categories using spatial-temporal words. [J] International Journal of Computer Vision 2008, 79:299 – 318

[5] Schuldt, C., Laptev, I., Caputo, B. Recognizing human actions: A local SVM approach. [C]//Proceedings of the Pattern Recognition, 17th International Conference Volume 3, IEEE Computer Society 2004:36

[6] Kovashka, A., Grauman, K. Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition. [C]// In: IEEE Conference on Computer Vision and Pattern Recognition. 2010

[7] Ando, R., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. [J] The Journal of Machine Learning Research 2005, 6:1817 – 1853

[8] Raina, R., Battle, A., Lee, H., Packer, B., Ng. A Self-taught learning: transfer learning from unlabeled data. [C]//Proceedings of the 24th international conference on Machine learning, ACM 2007:766

[9] Dollar, P., Rabaud, V., Cottrell, G., Belongie, S. Behavior recognition via sparse spatio-temporal features. [C]//Visual Surveillance and Performance Evaluation of Tracking and Surveillance 2nd Joint IEEE International Workshop 2005:65 – 72

[10] Laptev, I.: On space-time interest points. [J] International Journal of Computer Vision 64 2005:107 – 123

[11] Wong, S., Cipolla, R. Extracting spatiotemporal interest points using global information. [C]//Proc. of IEEE International Conference on Computer Vision. 2007

[12] Dikmen, M., Lin, D., Del Pozo, A., Cao, L., Fu, Y., Huang, T. A Study on Sampling Strategies in Space-Time Domain for Recognition Applications. Advances in Multimedia Modeling 2010: 465 – 476

[13] Wang, H., Ullah, M., Klaser, A., Laptev, I., Schmid, C. Evaluation of local spatiotemporal features for action recognition. [C]//British Machine Vision Conference.2009

[14] Grosse, R., Raina, R., Kwong, H., Ng, A.Shift-invariant sparse coding for audio classification. [J] cortex 2007: 8

[15] Mairal, J., Bach, F., Ponce, J., Sapiro, G. Online dictionary learning for sparse coding. [C]// Proceedings of the 26th Annual International Conference on Machine Learning, ACM 2009:689 – 696

[16] Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A. Discriminative learned dictionaries for local image analysis. [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2008: 1 – 8

[17] Yang, J., Yu, K., Gong, Y., Huang, T. Linear spatial pyramid matching using sparse coding for image classification. [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2009

[18] Yang, J., Yu, K., Huang, T. Supervised Translation-Invariant Sparse Coding. [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2010

[19] Olshausen, B., Field, D.Sparse coding with an overcomplete basis set: A strategy employed by V1? [J] Vision research 1997,(37):3311 – 3325

[20] Liu, Y., Cheng, J., Xu, C., Lu, H. Building topographic subspace model with transfer learning for sparse representation. [J] Neurocomputing 2010

[21] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T. Robust object recognition with cortex-like mechanisms. [J]PAMI 2007(29): 411 – 426

[22] Rodriguez, M.D., Ahmed, J., Shah, M. Action

mach: a spatio-temporal maximum average correlation height filter for action recognition.[C]//In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. 2008
[23] Jhuang, H., Serre, T.,Wolf, L., Poggio, T.A biologically inspired system for action recognition. [C]//Proc. of IEEE International Conference on Computer Vision. 2007
[24] Fathi, A., Mori, G.: Action recognition by learning mid-level motion features. [C]//IEEE Conference on Computer Vision and Pattern Recognition. 2008:1 – 8
[25] Bregonzio, M., Gong, S., Xiang, T. Recognising action as clouds of space-time interest points. [C]// IEEE Conference on Computer Vision and Pattern Recognition. 2009
[26] Yeffet, L., Wolf, L. Local Trinary Patterns for Human Action Recognition. [C]//Proc. of IEEE International Conference on Computer Vision. 200