Computer Science and Application 计算机科学与应用, 2018, 8(4), 455-463
Published Online April 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.84050

Hans汉斯

Activity Recognition of Middle-Level Attribution Based on Detachable Structure

Shaoyan Shi, Yongxuan Sun, Kewei Wu, Zhao Xie

School of Computer and Information, Hefei University of Technology, Hefei Anhui
Email: shaoyan_shi93@163.com, syx@hfut.edu.cn

Received: Apr. 2nd, 2018; accepted: Apr. 17th, 2018; published: Apr. 25th, 2018

Abstract

With the deeper research of activity recognition, more researchers pay attention to complex activity recognition. However, a large number of redundant information and noise reduce the accuracy of activity recognition. In order to solve this problem, the paper proposes temporally consistent middle attributes structure and matching patterns of activity by dynamic time warping algorithm. Through splitting videos and learning middle attribute of segments, we build the consistent temporal structure of each activity category effectively solving noise and redundancy. And it explains the process of activity from the perspective of timing and enhances the interpretability of activity analysis. The experiment results on Olympics dataset show that the algorithm can improve the accuracy of activity recognition effectively.

Keywords

Temporal Structure, Middle-Level Attribution, Template Matching, Activity Recognition

基于可拆分结构的中层属性行为识别

时少艳，孙永宣，吴克伟，谢　昭

合肥工业大学计算机与信息学院，安徽　合肥
Email: shaoyan_shi93@163.com, syx@hfut.edu.cn

收稿日期：2018年4月2日；录用日期：2018年4月17日；发布日期：2018年4月25日

摘　要

随着行为识别研究的深入，复杂行为识别受到了越来越多研究者的关注。然而复杂行为中存在的大量冗余信息以及噪声严重降低了行为识别的准确性。针对这一问题，本文提出了基于可拆分结构的复杂行为

文章引用: 时少艳, 孙永宣, 吴克伟, 谢昭. 基于可拆分结构的中层属性行为识别[J]. 计算机科学与应用, 2018, 8(4): 455-463. DOI: 10.12677/csa.2018.84050

中层属性时序一致性结构，并采用动态时间规整算法进行行为时序模式匹配。通过对视频行为时序进行拆分，学习行为的中层属性表达，构建行为中层属性的一致时序关系，有效去除了视频中的噪声和冗余信息，提高行为识别的准确率，并且能够从行为时序发展角度解释行为，增强了行为分析的可解释性。在Olympics数据集上的实验结果表明，该算法能够有效提高行为识别的准确率。

关键词

时序结构，中层属性，模板匹配，行为识别

1. 引言

行为分析是指从视频序列中抽取相关的视觉信息，采用合适的方法进行表达，通过解释这些视觉信息达到对人的行为模式进行分析和识别的目的，是计算机视觉研究领域备受关注并具挑战性的任务。现实世界中，由于人的行为具有多样性，复杂性等特点，通过计算机分析与理解人的行为，可以节约大量的人力物力消耗，对实现办公室或工业环境等各种重要场所的安全监控[1]，医疗看护[2]，人机交互系统[3]等任务也是非常重要的。

行为是指人们一切有目的的活动，它是由若干个基本时空要素构成的，具有明显的时空结构特性。近年来，一些研究者尝试通过建模行为的时序关系实现行为分析。Tang 等人[4]通过将视频序列映射到视觉特征空间，采用 HMM 模型捕获视频序列的时序结构。Bhattacharya 等人[5]提出谱模型建模复杂行为的运动属性动态变化，刻画了行为发展时序关系。以及 Ma 等人[6]通过假设行为正例得分函数及与行为负例得分函数差距随时间呈单调增加，采用 LSTM 损失单调变化模型惩罚违背时序单调性的观测输入，从而极大提升了行为分析的准确性。Liu 等人[7]构建了行为动作单元后缀树，对比行为基本动作单元的时序动态变化规律，从而提升复杂行为识别准确率。上述研究表明构建行为的时序关系有利于提升行为分析的性能。

时序关联体现了行为发展的基本过程，对其建模可以更有效的理解视频行为信息。大多数的研究只关注行为的整体时序结构，对视频整体序列进行建模。然而现实中，由于视频角度不同以及视频中存在大量无关目标或场景，对视频整体序列结构建模会包含与兴趣个体无关的大量冗余信息。有效剔除视频中的冗余信息，提取行为共性特征成为行为分析的关键问题。Grenander 在模式理论[8]中指出，通过要素之间的连接来重组模式概念，达到识别和分类模式的目的。Souza 等人[9]首次将模式理论应用到空间行为识别，通过空间行为基本要素结构拆分和重组去除冗余信息并实现视频动作解释。同样，针对复杂行为时序分析，可以通过行为基本时序要素结构拆分和重组过程剔除冗余信息，提取行为时序结构共性特征实现行为分析。

针对长时序视频中存在大量冗余和噪声信息干扰行为分析的问题，本文将 Grenander 模式拆分重组理论与行为时序分析相结合，将模式理论可拆分原理应用到行为分析时空领域，通过拆分行为以及重组行为时序中层属性，构建符合人类认知的一致行为中层属性时序关系模型，去除噪声以及冗余背景等信息干扰，提高长时序复杂行为识别性能。对比 Hilde Kuehne (2016) [12]提出的方法，本文主要贡献如下：

1) 首次将模式拆分重组理论与行为时序结构分析相结合，是模式理论继空间行为分析后的一次新的

尝试，通过在行为时序中层属性层面上进行筛选和重组，能够有效去除行为中的冗余和噪声，构建复杂行为的精简表达。

2) 采用无监督的方式学习行为中的中层属性表达，减少了手工标记数据的人工和时间消耗，扩大了复杂数据集的实际应用范围。

3) 采用动态时间规整算法对行为中层属性时序结构模式进行匹配，实现行为识别任务，并在 Olympics 数据集上验证了本文方法能够有效提升行为识别准确率。

2. 动作基元时序模型

不同于传统基于底层特征表达的行为识别，本文通过帧速度分割的方法对视频行为进行时序结构拆分，采用基于 LDA (Latent Dirichlet Allocation)的无监督方式学习视频分割段的中层属性表达，去除视频底层特征中的噪声和冗余信息，并通过重组过程构建行为的中层属性一致时序结构关系，进一步精简有效行为信息表达，提高行为识别的准确率。本文流程图如图 1 所示。

2.1. 构建动作基元表达

视频分割粒度决定了视频中层属性表达的准确性，若视频分割长度过小，中层属性则会包含较多噪声，反之，视频分割长度过大，中层属性就会很稀疏，削弱行为区别性，因此适宜的分割长度对于行为中层属性表达具有较大影响。Li 等人[10]研究表明，行为过程与行为速度变化密切相关，速度极值往往代表了行为变化关键点。如图 1 左图所示，本文采用帧速度的方法进行行为分割，帧速度曲线的极大值和极小值点分别表示动作的中心点和分割点。本文采用改进的 Harris 角点算法发现行为关键点，使用 Lucas-Kanade 光流得到关键点的轨迹，通过对每一帧跟踪关键点计算速度大小。

基于密集轨迹特征的行为表达已经被证明非常有效。因此我们采用密集轨迹特征对行为构建特征表达[11]。

行为分割段的中层属性反映了行为过程的简单动作，如打篮球行为中的跳跃，投球等简单动作，本文将分割段的中层属性称之为动作基元。给定训练视频集合 $V=\{\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_m,\},m\in[1,M]$，我们采用无监督的方式学习各个视频分割段的动作基元表达。对于一段视频 $\mathbf{v}$，我们首先在时序上将其分割为 R 个分割段，$\mathbf{v}=\{x_1,x_2,\cdots,x_R\}$。为了构建每个视频分割段的动作基元表达，我们首先采用 LDA (如图 2 所示)方法构建每个视频分割段的主题分布 $p(z|w,d)$。

给定训练视频的无序运动特征单词 $W=\{w_1,w_2,\cdots,w_L\},l\in[1,L]$，其中 l 为单词索引。训练视频的主题分布记作 $Z=\{z_1,z_2,\cdots,z_K\},k\in[1,K]$，$\phi$为主题的先验分布，则特征单词服从离散分布 $p(w|\phi,z)$。视频分割段的主题单词分布 $p(z|w,d)$ 通过采样混和运动模式分布得到，每种运动模式 z_k 即为对混合运动单词的采样得到。

对于每个视频分割段的主题分布 $p(z|w,d)$ 表示一种运动特性分布，则相似运动特性具有相似的主题分布，因此我们对所有视频分割段的主题分布进行相似性度量，根据相似程度，将所有视频分割段的主题分布划分为 N 簇，每一簇为一种动作基元，则所有训练视频动作基元集合为 $S=\{a_1,a_2,\cdots,a_n,\cdots,a_N\}$，每段视频 $\mathbf{v}=\{x_1,x_2,\cdots,x_R\}$ 对应的动作基元序列为 $\mathbf{a}=\{a_1,a_2,\cdots,a_R\}$，则通过分割以及属性学习，可以获得每段视频的初始动作基元时序结构。

2.2. 动作基元时序重组

为去除噪声和冗余，需要对初始动作基元时序结构进行预重组处理。

第一层：合并相邻相同动作基元；

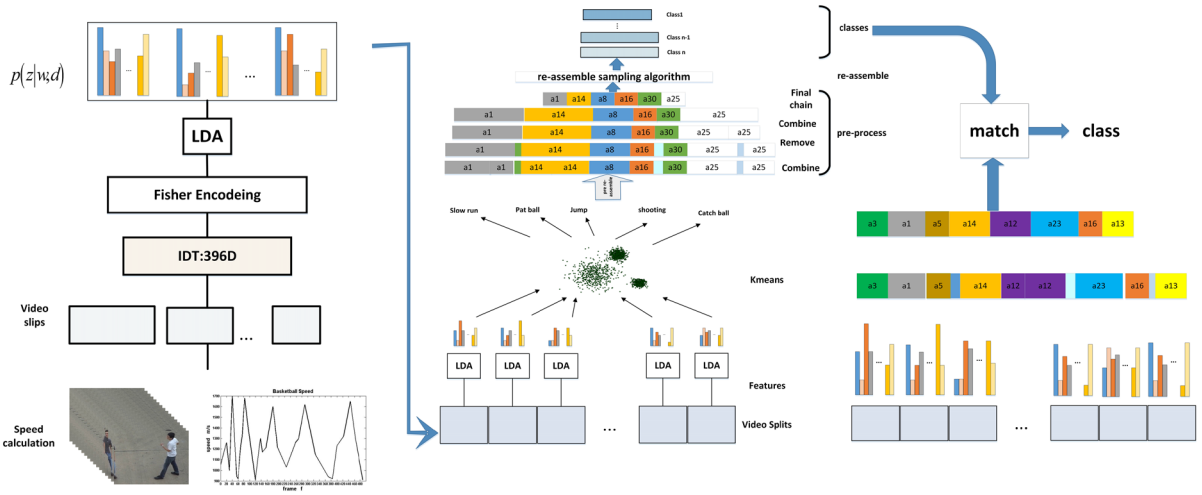

Figure 1. Pipeline of activity recognition of detachable structure
图 1. 可拆分结构行为识别流程图

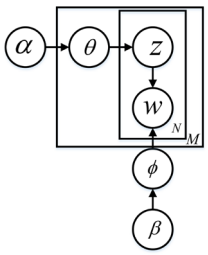

Figure 2. Model of LDA
图 2. LDA 模型图

第二层：剔除噪声，统计动作基元出现个数，将出现次数少于训练视频个数一半的基元剔除；

第三层：再次合并相邻相同基元。

对预处理后的时序结构进行随机采样和重组可能的动作基元一致时序结构，动作基元的连接强度(Strength)定义如下：

$$S\left(a_i,a_j\right)=\omega\cdot\tanh\left(\gamma\cdot f\right) \tag{1}$$

其中γ为动作基元共现频率，并且具有相同共现结构的时序段动作基元构成同一个候选空间。

$$\gamma=\sum_{\text{Dtraining}}\sum_{a_i,a_j\in v_n}\left[\!\left[a_i,a_j\right]\!\right] \tag{2}$$

当 a_i,a_j 共现时，$\left[\!\left[a_i,a_j\right]\!\right]=1$，否则为 0。$f$ 为动作基元的检测置信度，即主题分布相似概率，ω为时序连接因子，经验取值 $\omega=1.5$。则行为类别 c 重组后的时序结构概率为：

$$p(c)=\frac{1}{Z}\prod_{(a_i,a_j)\in c}S\left(a_i,a_j\right) \tag{3}$$

其中 Z 为归一化常量。

通过动作基元重组算法，可以获得每个类别的最优动作基元时序表达，采用动态时间规整算法将测试时序链与训练得到的时序链匹配完成行为识别任务。重组算法如算法 1 所示。

Algorithm 1. Actionlets re-assemble algorithm
算法 1. 动作基元序列重组

输入：预处理后包含 k_0 个动作基元的序列结构 c_0 及其结构概率 p_{c_0}，候选空间 $\mathbb{C}$，初始重组结构集合 $\Omega=\{c_0\}$，初始温度 T_0
输出：最优动作基元序列 c

1：初始化：$c=c_0$，$n=0$，$k_{\max}=k_0$
1：loop: $c\in\Omega$
3：for $i\leftarrow 1,k_{\max}$ do
4：for $j\leftarrow 1,3$ do
5：从第 i 个候选空间 $\mathbb{C}$ 选择第 j 个动作基元，并替换 c_0 中第 i 个时序段的动作基元，获得新的结构 c'，计算 c' 概率 $p_{c'}$，$z\sim U(0,1)$，$n=n+1$
6：if $p(c')>p(c)$ or $z<\exp\left(p(c')-p(c)\right)/T$
7：$c\leftarrow c'$，$\Omega\leftarrow c'$
8：else
9：$T\leftarrow T_0\times\alpha^n$
10：end if
11：end for
12：end for
13：比较重组结构集合Ω所有 p_c，选择概率最大的 c''，$c\leftarrow c''$
14：skip loop
15：比较重组结构集合Ω所有 p_c，选择概率最大的时序结构为最优结构 c

3. 实验结果评价

Olympics 数据集包含了运动员的各类运动视频。视频都是从 YouTube 上下载的，包含有 16 种运动类别，每个动作类别约 50 个视频，平均视频长度为 5~30 s。

在密集轨迹特征的提取中，本文采用原始算法参数设置，获得的 HOG (Histogram Of Gradient)、HOF (Histogram Of Flow)和 MBH (Motion Boundary Histograms)的维度分别是 96，108，96*2，HOG、HOF 和 MBH 级联构成密集轨迹特征，为 396 维。然后将密集轨迹特征降至 160 维，特征还原率为 0.97。随机采样 40 万特征进行 FV (Fisher Vector)特征参数计算，并将 FV 特征降至 100 维。在 LDA 主题模型构建单词码本的时候，我们从训练数据集中随机选择 1/2 的样本描述子进行 k-means 聚类，创建 5000 个单词。LDA 中的模型参数 α,β 分别设置为 0.05 和 1；在重组算法 1 中，经验上，我们设置初始 $T=2500$，$\alpha=0.9967$。

在图 3 中，通过对比重组前后识别结果可以看出，本文的重组算法可以有效提升识别准确率，重组过程中的筛选和合并过程对于噪声和冗余的去除是有效的。与同是采用密集轨迹特征并采用 FV 特征作为模型输入的 Hilde Kuehne (2016) [12]相比，我们探究了不同高斯混合模型聚类个数下的识别性能，从图 3 可以看出，本文方法在各高斯混合模型聚类个数下均优于 Hilde Kuehne (2016)，并且当聚类个数大于 128 时，Hilde Kuehne (2016)方法性能开始下降，本文方法的效果仍然很好。Hilde Kuehne (2016)采用 HMM 方法对所有视频片段分别学习得到其动作类别作为最终识别结果，忽略了视频中的冗余和噪声信息，因此是对所有视频数据建模，而本文方法首次采用无监督方式将视频底层信息映射至中层动作基元表达，能够提取更加精简的有用信息，弱化了噪声等信息干扰，并在重组过程中，将会对包含噪声和冗余信息的动作基元进行置信度筛选，极大降低了冗余和噪声信息干扰，因此本文方法优于 Hilde Kuehne (2016)方法。并且，Hilde Kuehne (2016)方法需要对所有训练片段分别学习动作表达分布，而本文方法可以对所有训练片段联合学习动作表达分布，如图 1 中间图所示，此过程能够有效节约训练样本单独的参数学习

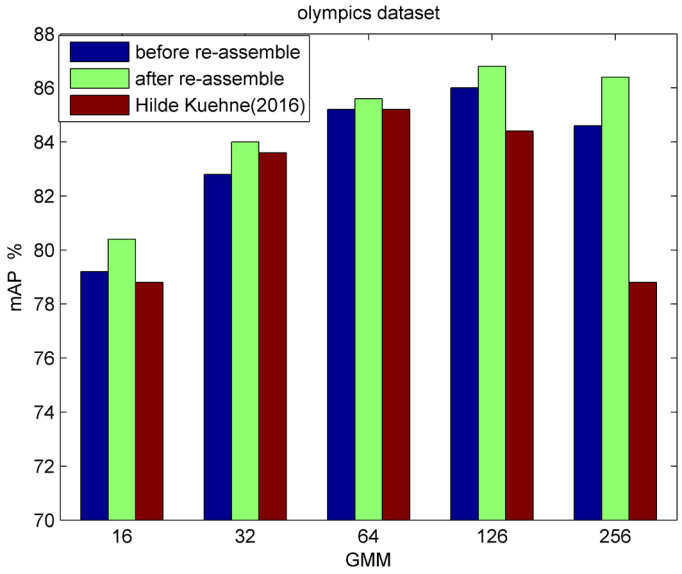

Figure 3. Comparison of recognition results on Olympics dataset
图 3. Olympics 数据集上识别结果对比

过程，训练相同的样本数目(70~80 样本数)，本文方法仅需 60 分钟左右，而 Hilde Kuehne (2016)方法需要 90 分钟左右，由此可见本文方法可以大大提高训练效率。

图 4 包含了 Olympics 数据集所有类别的 AP (Average Precision)识别结果。从图 4 中可以看出，与 HMM-FV [13]，TIAN LAN*[14]方法相比，我们的方法在绝大多数行为类别上都明显提高了识别性能。从图中可以看出，对于时序动作规律性较强的行为类别，如 basketball，bowling，high jump 等行为，我们的方法相比 HMM-FV，TIAN LAN*方法更能显著提高识别性能，并且与动作基元重组前行为识别结果相比，重组过程明显提高了识别性能。而对于一些语境环境较长，动作持续时间较短的快速动作，我们的方法准确率会有所下降，如 diving，hammer throw 等行为，这是由于对于背景较长，动作持续时间较短的快速动作，难以将其拆分为多个时序单元，导致多个很短时序动作基元级联为单个时序动作基元，重组过程中的基元剔除以及合并操作对于动作基元初始时序结构重组作用很小，由此导致整个重组过程对识别性能提高作用很小。

本文采用 LDA 模型学习分割段的主题分布，对所有分割段的主题分布采用 k-means 聚类为 K 类构建 K 种动作基元，并对不同主题个数 T 和主题聚类个数 K 的影响分别进行实验，从图 5 可以看出，在 Olympics 数据集中，当 $T=40, K=35$ 时，识别性能最好。

与近几年行为识别方法相比，如表 1 所示，可以看出本文方法效果明显高于其他方法。本文方法构建视频的动作基元时序结构，从行为演变过程角度分析行为，如图 6 所示，每个分割段大致对应于行为过程的一个简单动作，这些简单动作的时序连接构成复杂行为，说明本文方法能够有效在中高层语义上捕获行为运动过程，从行为认知过程上解释行为动态演变。

4. 结束语

由于复杂行为存在大量冗余信息和噪声，本文提出了构建复杂行为的动作基元表达，能够有效弱化噪声，并且通过学习行为共有的动作基元时序一致性结构关系，剔除了大量冗余和噪声信息。实验表明，本文提出的针对复杂行为进行拆分重组的精简行为表达能够有效提升行为识别的准确率以及训练速度。除此之外，对复杂行为的拆分重组思想对于未来越来越复杂的行为识别任务非常具有借鉴意义。

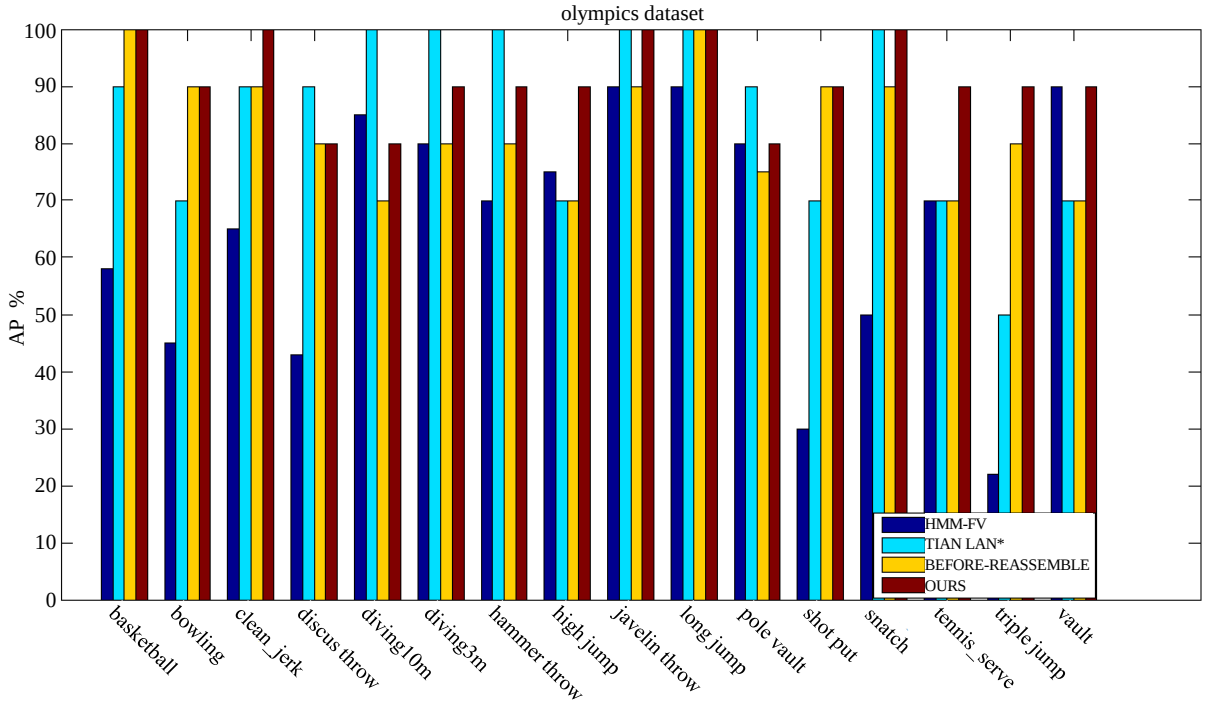

Figure 4. AP performance on Olympics dataset
图 4. Olympics 数据集上 AP 性能

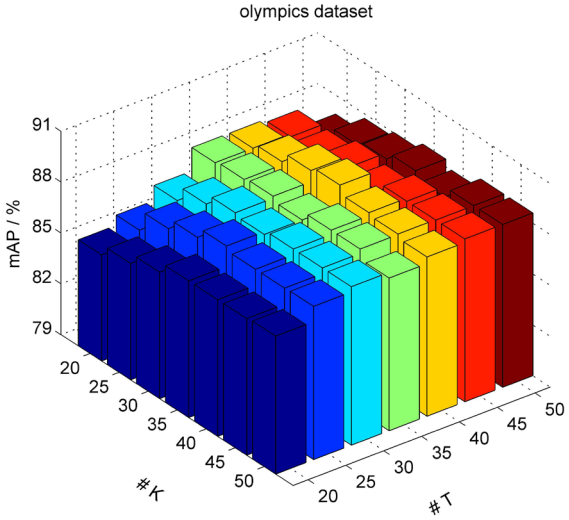

Figure 5. Influence of parameters of actionlets
图 5. 动作基元表达参数对性能影响

Table 1. Comparison of AP on Olympics dataset
表 1. Olympics 数据集上 mAP 性能比较

方法	mAP
Hilde Kuehne (2016) [12]	90.2%
Wang and Schmid (2013) [11]	90.1%
Jones and Shao (2014) [15]	74.6%
本文方法	90.5%

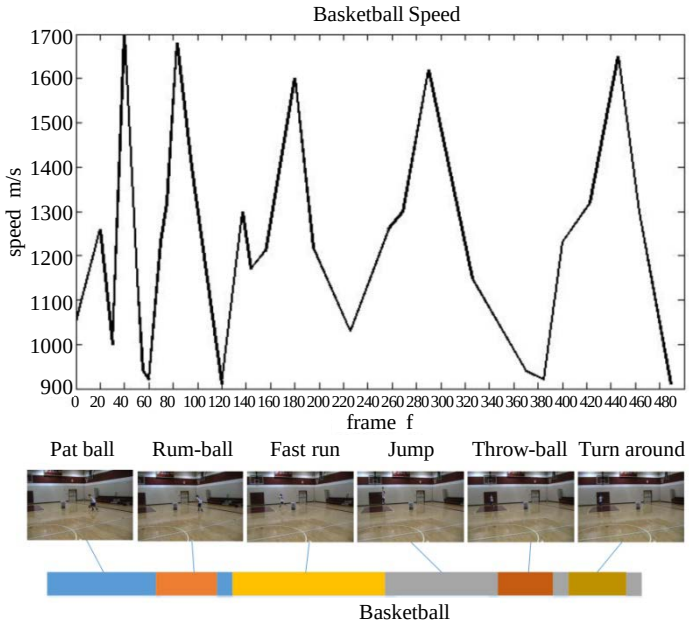

Figure 6. Relation between video segments and actionlets
图 6. 视频分割结果与动作基元对

致　谢

感谢国家重点研发计划(2017YFB1002203)和国家自然科学基金(No.61503111, No.61501467)的支持。

参考文献

[1] 雷庆, 陈锻生, 李绍滋. 复杂场景下的人体行为识别研究新进展[J]. 计算机科学, 2014, 41(12): 1-7.

[2] 胡琼, 秦磊, 黄庆明. 基于视觉的人体动作识别综述[J]. 计算机学报, 2013, 36(12): 2512-2524.

[3] Aggarwal, J.K. and Ryoo, M.S. (2011) Human Activity Analysis: A Review. *ACM Computing Surveys*, **43**, 1-43. https://doi.org/10.1145/1922649.1922653

[4] Tang, K., Li, F.F. and Koller, D. (2012) Learning Latent Temporal Structure for Complex Event Detection. *Computer Vision and Pattern Recognition*, IEEE, 1250-1257.

[5] Bhattacharya, S., Kalayeh, M.M., Sukthankar, R., *et al*. (2015) Recognition of Complex Events: Exploiting Temporal Dynamics between Underlying Concepts. *Computer Vision and Pattern Recognition*, 2243-2250.

[6] Ma, S., Sigal, L. and Sclaroff, S. (2016) Learning Activity Progression in LSTMs for Activity Detection and Early Detection. *Computer Vision and Pattern Recognition*, IEEE, 1942-1950.

[7] Liu, C., Wu, X. and Jia, Y. (2016) A Hierarchical Video Description for Complex Activity Understanding. *International Journal of Computer Vision*, **118**, 240-255. https://doi.org/10.1007/s11263-016-0897-2

[8] Grenander, U. (1993) General Pattern Theory: A Mathematical Study of Regular Structures. Clarendon Press, Oxford.

[9] Souza, F.D.M.D., Sarkar, S., Srivastava, A., *et al*. (2016) Spatially Coherent Interpretations of Videos Using Pattern Theory. *International Journal of Computer Vision*, **121**, 5-25. https://doi.org/10.1007/s11263-016-0913-6

[10] Li, K., Hu, J. and Fu, Y. (2012) Modeling Complex Temporal Composition of Actionlets for Activity Prediction. *European Conference on Computer Vision*, 286-299.

[11] Wang, H. and Schmid, C. (2014) Action Recognition with Improved Trajectories. *IEEE International Conference on Computer Vision*, IEEE, 3551-3558.

[12] Kuehne, H., Gall, J. and Serre, T. (2016) An End-to-End Generative Framework for Video Segmentation and Recognition. *Applications of Computer Vision*, IEEE, 1-8.

[13] Sun, C. and Nevatia, R. (2013) ACTIVE: Activity Concept Transitions in Video Event Classification. *IEEE International Conference on Computer Vision*, IEEE, 913-920.

[14] Lan, T., Zhu, Y., Zamir, A.R., *et al*. (2015) Action Recognition by Hierarchical Mid-Level Action Elements. *IEEE International Conference on Computer Vision*, IEEE, 4552-4560.

[15] Jones, S. and Shao, L. (2014) A Multigraph Representation for Improved Unsupervised/Semi-Supervised Learning of Human Actions. *Computer Vision and Pattern Recognition*, IEEE, 820-826.