J Real-Time Image Proc

实时图像处理

DOI 10.1007/s11554-013-0370-1

Doi 10.1007 / s11554-013-0370-1

O R I G I N A L R E S E A R C H P A P E R

O r i g i n a l r e s e a r h p a p e r

Real-time human action recognition based on depth motion maps

基于深度运动图的实时人体动作识别

Chen Chen • Kui Liu • Nasser Kehtarnavaz

陈陈•奎•刘•纳赛尔•凯特纳瓦兹

Received: 8 April 2013 / Accepted: 25 July 2013

收到:2013年4月8日/接受:2013年7月25日

Springer-Verlag Berlin Heidelberg 2013

Springer-VerlagBerlinHeidelberg2013年

Abstract This paper presents a human action recognition method by using depth motion maps (DMMs). Each depth frame in a depth video sequence is projected onto three orthogonal Cartesian planes. Under each projection view, the absolute difference between two consecutive projected maps is accumulated through an entire depth video sequence forming a DMM. An l2-regularized collaborative represen-tation classifier with a distance-weighted Tikhonov matrix is then employed for action recognition. The developed method is shown to be computationally efficient allowing it to run in real-time. The recognition results applied to the Microsoft Research Action3D dataset indicate superior performance of our method over the existing methods.

本文提出了一种基于深度运动图的人体动作识别方法。深度视频序列中的每个深度帧被投影到三个正交的笛卡尔平面上。在每个投影视图下，两个连续投影图之间的绝对差值通过一个完整的深度视频序列累积形成一个数字多用表。然后采用距离加权Tikhonov矩阵的l2正则化协同表示分类器进行动作识别。所开发的方法被证明是计算效率高，允许它在实时运行。应用于MicrosoftResearchAction3D数据集的识别结果表明，我们的方法比现有方法具有更好的性能。

Keywords Human action recognition Depth motion map RGBD camera Collaborative representation classifier

人体动作识别深度运动图RGBD摄像机协同表示分类器

1 Introduction

1引言

Human action recognition is an active research area in computer vision. Earlier attempts at action recognition have involved using video sequences captured by video cameras. Spatio-temporal features are widely used for recognizing human actions, e.g. [1–[6](#page8)]. As imaging tech-nology advances, it has become possible to capture depth information in real-time. Compared with conventional images, depth maps are insensitive to changes in lighting conditions and can provide 3D information toward distin-guishing actions that are difficult to characterize using

人体动作识别是计算机视觉领域中一个非常活跃的研究领域。早期的动作识别尝试包括使用摄像机拍摄的视频序列。时空特征被广泛用于识别人类行为，例如[1-]。随着成像技术的进步，实时获取深度信息已成为可能。与传统图像相比，深度图对光照条件的变化不敏感，可以提供三维信息，用于描述难以使用的抛光行为

C. Chen (&) K. Liu N. Kehtarnavaz

陈(及)刘锦洪

The University of Texas at Dallas, Richardson, TX, USA e-mail: chenchen870713@gmail.com

德克萨斯大学达拉斯，理查森，TX电子邮件:chen870713@gmail.com

conventional images. Figure [1](#page2) shows two examples con-sisting of nine depth maps of the action Golf swing and the action Forward kick. Since the release of low cost depth sensors, in particular Microsoft Kinect and ASUS Xtion, many research works have been carried out on human action recognition using depth imagery, e.g. [[7](#page8)–[13](#page8)]. As noted in [[14](#page9)], 3D joint positions of a person’s skeleton estimated from depth images provide additional informa-tion to achieve action recognition.

常规图像。图显示了两个示例，其中包含了动作高尔夫挥杆和动作向前踢的九个深度图。自从发布了低成本的深度传感器，特别是微软Kinect和ASUSXtion，许多研究工作已经开始使用深度图像进行人体动作识别，例如[-]。正如在[]中提到的，通过深度图像估计的人体骨骼的3D关节位置提供了额外的信息来实现动作识别。

In this paper, the problem of human action recognition from depth map sequences is examined from the perspective of computational efficiency. These images are captured by an RGBD camera. Specifically, the depth motion maps (DMMs) generated by accumulating motion energy of projected depth maps in three projective views (front view, side view, and top view) are used as feature descriptors. Compared with 3D depth maps, DMMs are 2D images that provide an encoding of motion characteristics of an action. Motivated by the suc-cess of sparse representation in face recognition [[15](#page9)–[18](#page9)] and image classification [[18](#page9), [19](#page9)], an l2-regularized collaborative representation classifier is utilized which seeks a match of an unknown sample via a linear combination of training samples from all the classes. The class label is then derived according to the class which best approximates the unknown sample. Basically, our introduced method involves a spatio-temporal motion representation based on DMMs followed by an l2-regularized collaborative representation classifier with a dis-tance-weighted Tikhonov matrix to perform computationally efficient action recognition.

本文从计算效率的角度研究了深度图序列中的人类行为识别问题。这些图像由一个RGBD相机捕捉。特别地，通过在三个投影视图(正视图、侧视图和俯视图)中积累投影深度图的运动能量，生成深度运动图(dmm)作为特征描述子。与3D深度图相比，dmm是提供动作运动特征编码的2D图像。基于稀疏表示在人脸识别和图像分类中的成功应用，提出了一种l2正则化的协同表示分类器，该分类器通过从所有类别的训练样本中提取一个线性组合来寻找未知样本的匹配。然后根据最接近未知样本的类派生类标签。基本上，我们所介绍的方法包括一个基于DMMs的时空运动表示，然后是一个带有离散加权Tikhonov矩阵的l2正则化协同表示分类器，以执行计算效率高的动作识别。

The rest of the paper is organized as follows. In Sect. [2](#page2), related works are presented. In Sect. [3](#page2), the details of generating DMMs feature descriptors are stated. In Sect. [4](#page3), the sparse representation classifier (SRC) is first introduced and then the l2-regularized collaborative representation classifier is described for performing action recognition.

文章的其余部分如下。宗。，介绍相关工作。宗。详细阐述了数字图像处理系统特征描述符的生成方法。宗。首先介绍了稀疏表示分类器(SRC)，然后描述了用于动作识别的l2正则化协同表示分类器。

123

123

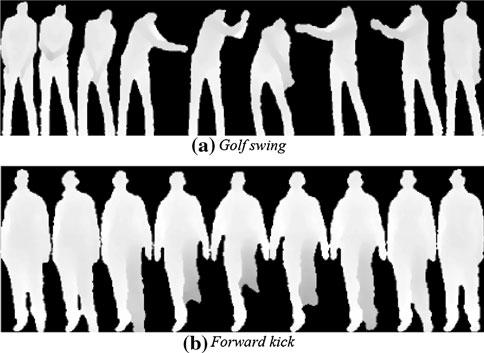


J Real-Time Image Proc

实时图像处理

Fig. 1 Examples of depth map sequences for a Golf swing action, and b Forward kick action

图1高尔夫挥杆动作和前踢动作的深度图序列示例



The experimental results are reported in Sect. 5. Finally, in Sect. [6](#page8), concluding remarks are stated.

第五节介绍了实验结果。最后，在第五节中作了总结。

2 Related works

2相关工程

Space–time based methods such as space–time volumes, spatio-temporal features, and trajectories have been widely utilized for human action recognition from video sequences captured by traditional RGB cameras. In [[1](#page8)], spatio-temporal interest points coupled with an SVM classifier was used to achieve human action recognition. Cuboid descriptors were employed in [[2](#page8)] for action representation. In [[3](#page8)], SIFT-feature trajectories modeled in a hierarchy of three abstraction levels were used to recognize actions in video sequences. Various local motion features were gathered as spatio-temporal bag-of-features (BoF) in [[4](#page8)] to perform action classification. Motion-energy images (MEI) and motion-history images (MHI) were introduced in [[5](#page8)] as motion templates to model spatial and temporal characteristics of human actions in videos. In [[6](#page8)], a hierarchical extension for computing dense motion flow from MHI was presented. A major shortcoming associated with using these intensity-based or color-based methods is the sensitivity of recognition to illumination variations, limiting the recognition robustness.

基于时空体积、时空特征和轨迹等方法已广泛应用于传统RGB摄像机采集的视频序列中的人体行为识别。在[]中，利用时空兴趣点与支持向量机分类器相结合的方法来实现人的行为识别。长方体描述符用于动作表示[]。在[]中，sift特征轨迹模型在三个抽象层次的层次结构中被用来识别视频序列中的动作。在[]中，各种局部运动特征被收集为时空特征袋(BoF)来执行动作分类。运动能量图像(MEI)和运动历史图像(MHI)作为运动模板，以模型的时空特征，人的行为在视频。在[]中，提出了一个计算密集运动流的层次扩展MHI。使用这些基于亮度或颜色的方法的主要缺点是识别对光照变化的敏感性，限制了识别的鲁棒性。

With the release of RGBD sensors, research into action recognition based on depth information has grown. Skele-ton-based approaches utilize locations of skeletal joints extracted from depth images. In [[7](#page8)], a view invariant posture representation was devised using histograms of 3D joint locations (HOJ3D) within a modified spherical coor-dinate system. HOJ3D were re-projected using LDA and clustered into k posture visual words. The temporal evo-lutions of these visual words were modeled by a discrete hidden Markov model. In [[8](#page8)], a Naive-Bayes-Nearest-Neighbor (NBNN) classifier was employed to recognize

随着rbd传感器的发布，基于深度信息的动作识别研究得到了发展。基于骨架吨的方法利用从深度图像骨骼关节的位置提取。在[]中，利用改进的球面坐标系中三维关节位置(HOJ3D)的直方图设计了一种视点不变的姿态表示。使用LDA重新投影HOJ3D并聚类成k姿态视觉词。这些视觉词汇的时间演化由一个离散的隐马尔可夫模型模型来模拟。在[]中，使用Naive-Bayes-Nearest-Neighbor(NBNN)分类器进行识别

human actions based on Eigen Joints (i.e., position differ-ences of joints) combining static posture, motion, and offset information. Such skeleton-based approaches have limitations due to inaccuracies in skeletal estimation. Moreover, the skeleton information is not always available in many applications.

基于本征关节(即关节的位置不同)的人类动作，结合静态姿势、运动和偏移信息。由于骨架估计的不准确性，这种基于骨架的方法有其局限性。此外，在许多应用程序中，骨架信息并不总是可用的。

There are methods that involve extracting spatio-tem-poral features from the entire set of points in a depth map sequence to distinguish different actions. An action graph was employed in [[9](#page8)] to model the dynamics of actions and a collection of 3D points were used to characterize pos-tures. However, the 3D points sampling scheme used generated a large amount of data leading to a computa-tionally expensive training step. In [[10](#page8)], a DMM-based histogram of oriented gradients (HOG) was utilized to compactly represent the body shape and movement infor-mation toward distinguishing actions. In [[11](#page8)], random occupancy pattern (ROP) features were extracted from depth images using a weighted sampling scheme. A sparse coding approach was utilized to robustly encode ROP features for action recognition and the features were shown to be robust to occlusion. In [[12](#page8)], 4D space–time occu-pancy patterns were used as features which preserved spatial and temporal contextual information coping with intra-class variations. A simple classifier based on the cosine distance was then used for action recognition.

有些方法涉及从深度图序列中的整个点集中提取空间-时间-时间特征，以区分不同的动作。在[]中使用了一个动作图来建立动作的动力学模型，并使用了一组3D点来描述动作。然而，使用的三维点采样方案产生了大量的数据，导致了计算代价昂贵的训练步骤。在[]中，一个基于dmm的方向梯度直方图(HOG)被用来紧密地代表身体形状和朝着区分动作的运动信息。在[]，随机占用模式(ROP)特征提取深度图像使用加权采样方案。采用稀疏编码的方法对ROP特征进行稳健编码以进行动作识别，结果表明该特征对遮挡具有较强的鲁棒性。在文献[]中，我们使用了四维时空概率模式作为特征，保留了空间和时间上下文信息以应对类内变化。然后采用基于余弦距离的简单分类器进行动作识别。

In [[13](#page8)], a hybrid solution combining skeleton and depth information was used for action recognition. 3D joint position and local occupancy patterns were used as fea-tures. An actionlet ensemble model was then learnt to represent each action and to capture intra-class variations.

在[]中，一种结合骨架和深度信息的混合解决方案被用于动作识别。以三维关节位置和局部占位模式为特征。然后学习一个actionlet集成模型来表示每个动作并捕捉类内的变化。

In general, the above references do not elaborate on the computational complexity aspect of their solutions and do not provide actual real-time processing times. In contrast to the existing methods, in this work, both the computational complexity and the processing times associated with each component of our method are reported.

一般来说，上述文献没有详细说明其解的计算复杂性，也没有提供实际的实时处理时间。与现有的方法相比，本文报告了计算复杂度和与每个部分相关的处理时间。

3 Depth motion maps as features

3深度运动图作为特征

A depth map can be used to capture the 3D structure and shape information. Yang et al. [[10](#page8)] proposed to project depth frames onto three orthogonal Cartesian planes for the purpose of characterizing the motion of an action. Due to its computational simplicity, the same approach in [[10](#page8)] is adopted in this work while the procedure to obtain DMMs is modified. More specifically, each 3D depth frame is used to generate three 2D projected maps corresponding to front, side, and top views, denoted by mapv where v 2 f f ; s; tg. For a point ðx; y; zÞ in a depth frame with z denoting the depth value in a right-handed coordinate system, the pixel value in three projected maps is indicated by z, x, and y,

深度图可以用来捕捉三维结构和形状信息。杨等人[]提出投影深度框架到三个正交笛卡尔平面，目的是刻画运动的一个行动。由于计算简单，本文采用了文献[]中的方法，同时修改了获取dmm的步骤。更具体地说，每个三维深度框架用于生成三个二维投影图，分别对应于前视图、侧视图和顶视图，用mapv表示，其中v2f;s;tg。对于深度帧中的点x;y;z，z表示右手坐标系中的深度值，三个投影图中的像素值由z、x和y表示,

123

123

J Real-Time Image Proc

实时图像处理

respectively. Different from [10], for each projected map, the motion energy is calculated here as the absolute dif-ference between two consecutive maps without threshold-ing. For a depth video sequence with N frames, DMMv is obtained by stacking the motion energy across an entire depth video sequence as follows:

分别。与[10]不同的是，对于每个投影映射，运动能量在这里被计算为两个不带阈值的连续映射之间的绝对差值。对于n帧的深度视频序列，通过将整个深度视频序列的运动能量叠加得到DMMv，如下所示:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ¼ | b |  |  |  | ð |  | Þ |  |
| DMMv |  | mapvi | 1 |  |
|  | X |  | mapvi 1 ; |  |  |  |
|  |  | i¼a |  |  |  |  |  |  |  |
| where i represents the frame index; mapi | | | | | | is the projected | | |  |
|  |  |  |  |  | v |  |  |  |  |

map of ith frame under projection view v; a 2 f2; . . .; Ng and b 2 f2; . . .; Ng denote the starting frame and the end frame index. It should be noted that not all the frames in a depth video sequence are used to generate DMMs. This point is discussed further in the experimental setup section. A bounding box is then set to extract the non-zero region as the foreground in each DMM.

投影视图下的第i帧图v;a2f2;...;Ng和b2f2;...;Ng表示起始帧和终止帧索引。需要注意的是，并非所有深度视频序列中的帧都用于生成dmm。这一点将在实验装置部分进一步讨论。然后设置一个包围盒来提取非零区域作为每个数字多用表的前景。

Let the foreground extracted DMM be denoted by DMMv hereafter. Two examples of DMMv generated from the Tennis serve and Forward kick video sequences are shown in Fig. [2](#page3). DMMs from the three projection views effectively capture the characteristics of the motion in a distinguishable way. That is the reason here for using DMMs as feature descriptors for action recognition. Since DMMv of different action video sequences may have different sizes, bicubic interpolation is used to resize all DMMv under the same projection view to a fixed size in order to reduce the intra-class variability, for example due to different subject heights. The size of DMMf is mf nf , the size of DMMs is ms ns, and the size of DMMt is mt nt. Since pixel values are used as features, they are normalized between 0 and 1 to avoid large pixel values dominating the feature set. The resized and normalized DMM is denoted by DMMv. For an action video

前景提取的DMM以后用DMMv表示。两个例子的DMMv生成的网球发球和前进踢视频序列如图所示。.来自三个投影视图的dmm以可分辨的方式有效地捕获了运动的特征。这就是使用dmm作为动作识别的特征描述符的原因。由于不同动作视频序列的DMMv可能具有不同的大小，因此我们使用双三次插值来将同一投影视图下的所有DMMv调整为固定大小，以减少类内的变化，例如由于主题高度不同而引起的变化。Dmmf的尺寸为mfnf，DMMs的尺寸为msns，DMMt的尺寸为mt。由于像素值被用作特征，它们被标准化在0到1之间，以避免大的像素值支配特征集。调整大小和规范化的DMM由DMMv表示。对于一个动作视频

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| sequence | | with three DMMs, | | | | a feature | | | vector of size |  |
| mf nf | | þ ms ns þ mt nt | | | | 1 is thus formed to be h ¼ | | | |  |
| vec |  | | f ; vec |  | s ; vec | |  | t T |  |  |
| DMM | | DMM | DMM | by concatenat- |  |

ing the three vectorized DMMs; vecð Þ indicates the vec-torization operator and T the matrix transpose. The feature vector encodes the 4D characteristics of an action video sequence. Note that the HOG descriptors of the DMMs are not computed here as done in [[10](#page8)] and image resizing is applied to DMMs but not to each projected depth map as done in [[10](#page8)]. As a result, the computational complexity of the feature extraction process is greatly reduced.

使用三个向量化DMMs;vec表示向量化算子，t表示矩阵转置。特征矢量对动作视频序列的四维特征进行编码。注意，dmm的HOG描述符在这里不像在[]中那样计算，图像调整应用于dmm，而不是像在[]中那样应用于每个投影深度映射。从而大大降低了特征提取过程的计算复杂度。

4 l2-regularized collaborative representation classifier

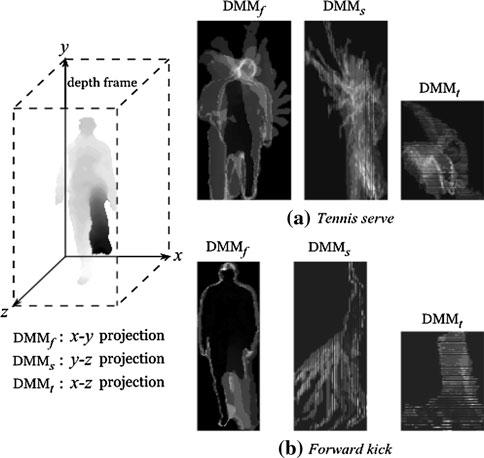
4个l2正则化协同表示分类器

Sparse representation (or sparse coding) has been an active research area in the machine learning community due to its success in face recognition [[15](#page9)–[18](#page9)] and image classification [[18](#page9), [19](#page9)]. The central idea of the SRC is to represent a test

稀疏表示(或稀疏编码)已成为一个活跃的研究领域，在机器学习社区由于其成功的人脸识别[-]和图像分类[，]。Src的中心思想是代表一个测试

Fig. 2 DMMv generated from a Tennis serve, and b Forward kick depth action video sequences

图2dmmv由网球发球产生，b前踢深度动作视频序列



sample according to a small number of atoms sparsely chosen out of an over-complete dictionary formed by all the available training samples. Consider a dataset with C classes of training samples arranged column-wise A ¼ ½A1; A2; . . .; AC & 2 Rd n, where Ajðj ¼ 1; . . .; CÞ is the subset of the training samples associated with class j, d is the dimension of training samples and n is the total number of training samples from all the classes. A test sample g 2 Rd can be represented as a sparse linear combination of the training samples, which can be formulated as,

样本根据少量的原子疏选出一个超完整的由所有可用的训练样本组成的词典。考虑一个有c类训练样本的数据集，按列排列a112A1;A2;...;AC&2Rdn，其中ajj11...;...;c是与类j相关联的训练样本的子集，d是训练样本的维数，n是所有类的训练样本的总数。一个测试样本g2Rd可以表示为训练样本的稀疏线性组合，它可以表示为:

g ¼ Aa; ð2Þ

G 1 / 4 Aa; 2

where a ¼ ½ a1; a2; . . .; aC& is an n 1 vector of coefficients corresponding to all the training samples and ajðj ¼ 1; . . .; CÞ denotes the subset of the coefficients associated with the training samples from the jth class, i.e. Aj. From a practical standpoint, one cannot directly solve for a since ([2](#page3)) is typically under-determined [[17](#page9)]. To reach a solution, one can solve the following l1 norm minimization problem,

其中a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a版本a。从实用的角度来看，一个人不能直接解决一个问题，因为()通常是不确定的[]。要得到一个解，可以解决下面的l1范数最小化问题,

* o

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| a^ ¼ arg | min | k | g |  | A | 2 | þhk a k1 | ; | ð3Þ |  |
| a |  |  | ak2 |  |

where h is a scalar regularization parameter which balances the influence of the residual and the sparsity term. The class label of g is then obtained via, where ej ¼ g Aja^j 2. The reader is referred to [] for more details.

其中h是一个标量正则化参数，平衡了剩余项和稀疏项的影响。然后通过获得g的类标号，其中ej1--4gAja^j2。要了解更多细节，请向读者查询。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ð | g | Þ ¼ | j | ej |  | ð |  | Þ |  |
| class |  | arg min |  | 4 |  |  |

As described in [[20](#page9)], it is the collaborative representation, i.e. use of all the training samples as a dictionary, but not the

正如在[]中所描述的，它是协作表示，即将所有的训练样本用作字典，而不是

123

123

J Real-Time Image Proc

实时图像处理

l1-norm sparsity constraint, that improves the classification accuracy. The l2-regularized approach generates comparable results but with significantly lower computational com-plexity [20, [21](#page9)]. Therefore, here the l2-regularized approach is used for action recognition. As mentioned in Sect. [3](#page2), each

L1-范数稀疏约束，提高了分类精度。L2正则化方法产生了可比较的结果，但计算复杂度明显较低[20，]。因此，本文采用l2正则化方法进行动作识别。正如在第节中提到的。登革热专页、每页