Image and Vision Computing 28 (2010) 976–990

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

A survey on vision-based human action recognition

Ronald Poppe *

Human Media Interaction Group, Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands

ARTICLE INFO

Article history:
Received 20 February 2009
Received in revised form 1 September 2009
Accepted 30 November 2009

Keywords:
Human action recognition
Motion analysis
Action detection

ABSTRACT

Vision-based human action recognition is the process of labeling image sequences with action labels. Robust solutions to this problem have applications in domains such as visual surveillance, video retrieval and human–computer interaction. The task is challenging due to variations in motion performance, recording settings and inter-personal differences. In this survey, we explicitly address these challenges. We provide a detailed overview of current advances in the field. Image representations and the subsequent classification process are discussed separately to focus on the novelties of recent research. Moreover, we discuss limitations of the state of the art and outline promising directions of research.

1. Introduction

We consider the task of labeling videos containing human motion with action classes. The interest in the topic is motivated by the promise of many applications, both offline and online. Automatic annotation of video enables more efficient searching, for example finding tackles in soccer matches, handshakes in news footage or typical dance moves in music videos. Online processing allows for automatic surveillance, for example in shopping malls, but also in smart homes for the elderly to support aging in place. Interactive applications, for example in human–computer interaction or games, also benefit from the advances in automatic human action recognition.

In this section, we first discuss related surveys and present the scope of this overview. Also, we outline the main characteristics and challenges of the field as these motivate the various approaches that are reported in literature. Finally, we briefly describe the most common datasets. In its simplest form, vision-based human action recognition can be regarded as a combination of feature extraction, and subsequent classification of these image representations. We discuss these two tasks in Sections 2 and 3, respectively. While many works will be described and analyzed in more detail, we do not intend to give complete coverage of all works in the area. In Section 4, we discuss limitations of the state of the art and outline future directions to address these.

1.1. Scope of this overview

The area of human action recognition is closely related to other lines of research that analyze human motion from images and video. The recognition of movement can be performed at various levels of abstraction. Different taxonomies have been proposed and here we adopt the hierarchy used by Moeslund et al. [90]: action primitive, action and activity. An action primitive is an atomic movement that can be described at the limb level. An action consists of action primitives and describes a, possibly cyclic, whole-body movement. Finally, activities contain a number of subsequent actions, and give an interpretation of the movement that is being performed. For example, “left leg forward” is an action primitive, whereas “running” is an action. “Jumping hurdles” is an activity that contains starting, jumping and running actions.

We focus on actions and do not explicitly consider context such as the environment (e.g. [119]), interactions between persons (e.g. [105,122]) or objects (e.g. [47,91]). Moreover, we consider only full-body movements, which excludes the work on gesture recognition (see [30,89]).

In the field of gait recognition, the focus is on identifying personal styles of walking movement, to be used as a biometric cue. The aim of human action recognition is opposite: to generalize over these variations. This is an arbitrary process as there is often significant intra-class variation. Recently, there have been several approaches that aim at simultaneous recognition of both action, and style (e.g. [22,28,152]). In this overview, we will discuss mainly those approaches that can deal with a variety of actions.

1.2. Surveys and taxonomies

There are several existing surveys within the area of vision-based human motion analysis and recognition. Recent overviews by Forsyth et al. [38] and Poppe [109] focus on the recovery of human poses and motion from image sequences. This can be regarded as a regression problem, whereas human action recognition is a

* Tel.: +31 534892836.
E-mail address: poppe@ewi.utwente.nl

0262-8856/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.imavis.2009.11.014

classification problem. Nevertheless, the two topics share many similarities, especially at the level of image representation. Also related is the work on human or pedestrian detection (e.g. [29,41,43]), where the task is to localize persons within the image.

Broader surveys that cover the above mentioned topics, including human action recognition, appear in [2,10,42,69,90,143,153]. Bobick [10] uses a taxonomy of movement recognition, activity recognition and action recognition. These three classes correspond roughly with low-level, mid-level and high-level vision tasks. It should be noted that we use a different definition of action and activity. Aggarwal and Cai [2], and later Wang et al. [153], discuss body structure analysis, tracking and recognition. Gavrila [42] uses a taxonomy of 2D approaches, 3D approaches and recognition. Moeslund et al. [90] use a functional taxonomy with subsequent phases: initialization, tracking, pose estimation and recognition. Within the recognition task, scene interpretation, holistic approaches, body-part approaches and action primitives are discussed. A recent survey by Turaga et al. [143] focuses on the higher-level recognition of human activity. Krüger et al. [69] additionally discuss intention recognition and imitation learning.

We limit our focus to vision-based human action recognition to address the characteristics that are typical for the domain. We discuss image representation and action classification separately as these are the two parts that are present in every action recognition approach. Due to the large variation in datasets and evaluation practice, we discuss action recognition approaches conceptually, without presenting detailed results. We focus on recent work, which has not been discussed in previous surveys. In addition, we present a discussion that focuses on promising work and points out future directions.

1.3. Challenges and characteristics of the domain

In human action recognition, the common approach is to extract image features from the video and to issue a corresponding action class label. The classification algorithm is usually learned from training data. In this section, we discuss the challenges that influence the choice of image representation and classification algorithm.

1.3.1. Intra- and inter-class variations

For many actions, there are large variations in performance. For example, walking movements can differ in speed and stride length. Also, there are anthropometric differences between individuals. Similar observations can be made for other actions, especially for non-cyclic actions or actions that are adapted to the environment (e.g. avoiding obstacles while walking, or pointing towards a certain location). A good human action recognition approach should be able to generalize over variations within one class and distinguish between actions of different classes. For increasing numbers of action classes, this will be more challenging as the overlap between classes will be higher. In some domains, a distribution over class labels might be a suitable alternative.

1.3.2. Environment and recording settings

The environment in which the action performance takes place is an important source of variation in the recording. Person localization might prove harder in cluttered or dynamic environments. Moreover, parts of the person might be occluded in the recording. Lighting conditions can further influence the appearance of the person.

The same action, observed from different viewpoints, can lead to very different image observations. Assuming a known camera viewpoint restricts the use to static cameras. When multiple cameras are used, viewpoint problems and issues with occlusion can be alleviated, especially when observations from multiple views can be combined into a consistent representation. Dynamic backgrounds increase the complexity of localizing the person in the image and robustly observing the motion. When using a moving camera, these challenges become even harder. In vision-based human action recognition, all these issues should be addressed explicitly.

1.3.3. Temporal variations

Often, actions are assumed to be readily segmented in time. Such an assumption moves the burden of the segmentation from the recognition task, but requires a separate segmentation process to have been employed previously. This might not always be realistic. Recent work on action detection (see Section 3.3) addresses this issue.

Also, there can be substantial variation in the rate of performance of an action. The rate at which the action is recorded has an important effect on the temporal extent of an action, especially when motion features are used. A robust human action recognition algorithm should be invariant to different rates of execution.

1.3.4. Obtaining and labeling training data

Many works described in this survey use publicly available datasets hat are specifically recorded for training and evaluation. This provides a sound mechanism for comparison but the sets often lack some of the earlier mentioned variations. Recently, more realistic datasets have been introduced (see also Section 1.4). These contain labeled sequences gathered from movies or web videos. While these sets address common variations, they are still limited in the number of training and test sequences.

Also, labeling these sequences is challenging. Several automatic approaches have been proposed, for example using web image search results [55], video subtitles [48] and subtitle to movie script matching [20,26,73]. Gaidon et al. [40] present an approach to rerank automatically extracted and aligned movie samples but manual verification is usually necessary. Also, performance of an action might be perceived differently. A small-scale experiment showed significant disagreement between human labeling and the assumed ground-truth on a common dataset [106]. When no labels are available, an unsupervised approach needs to be pursued but there is no guarantee that the discovered classes are semantically meaningful.

1.4. Common datasets

The use of publicly available datasets allows for the comparison of different approaches and gives insight into the (in)abilities of respective methods. We discuss the most widely used sets.

1.4.1. KTH human motion dataset

The KTH human motion dataset (Fig. 1a [125]) contains six actions (walking, jogging, running, boxing, hand waving and hand clapping), performed by 25 different actors. Four different scenarios are used: outdoors, outdoors with zooming, outdoors with different clothing and indoors. There is considerable variation in the performance and duration, and somewhat in the viewpoint. The backgrounds are relatively static. Apart from the zooming scenario, there is only slight camera movement.

1.4.2. Weizmann human action dataset

The human action dataset (Fig. 1b [9]) recorded at the Weizmann institute contains 10 actions (walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, jumping jack and skip), each performed by 10 persons. The backgrounds are static and foreground silhouettes are included in the dataset. The viewpoint is static. In addition to this dataset, two separate sets of sequences were recorded for robustness evaluation. One set shows

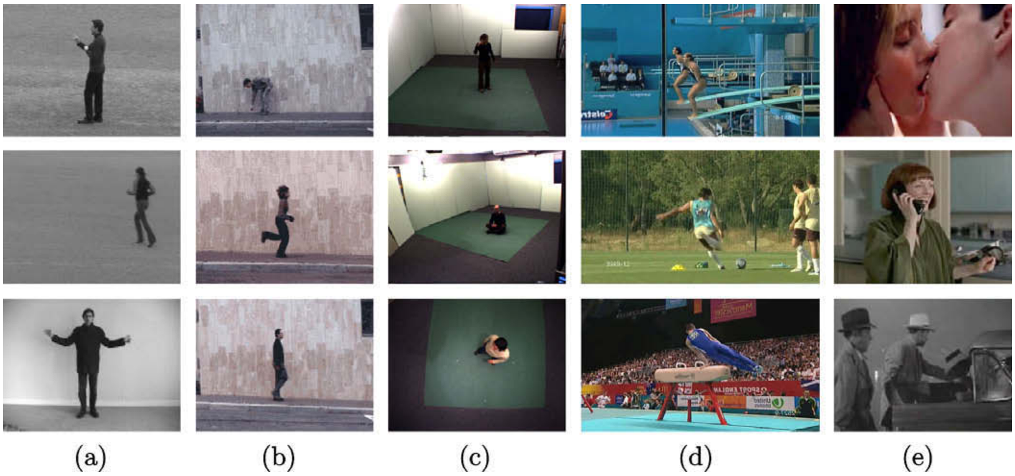

Fig. 1. Example frames of (a) KTH dataset, (b) Weizmann dataset, (c) Inria XMAS dataset, (d) UCF sports action dataset and (e) Hollywood human action dataset.

walking movement viewed from different angles. The second set shows fronto-parallel walking actions with slight variations (carrying objects, different clothing, different styles).

1.4.3. INRIA XMAS multi-view dataset

Weinland et al. [166] introduced the IXMAS dataset (Fig. 1c) that contains actions captured from five viewpoints. A total of 11 persons perform 14 actions (check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point, pick up, throw over head and throw from bottom up). The actions are performed in an arbitrary direction with regard to the camera setup. The camera views are fixed, with a static background and illumination settings. Silhouettes and volumetric voxel representations are part of the dataset.

1.4.4. UCF sports action dataset

The UCF sports action dataset (Fig. 1d [120]) contains 150 sequences of sport motions (diving, golf swinging, kicking, weightlifting, horseback riding, running, skating, swinging a baseball bat and walking). Bounding boxes of the human figure are provided with the dataset. For most action classes, there is considerable variation in action performance, human appearance, camera movement, viewpoint, illumination and background.

1.4.5. Hollywood human action dataset

The Hollywood human action dataset (Fig. 1e [73]) contains eight actions (answer phone, get out of car, handshake, hug, kiss, sit down, sit up and stand up), extracted from movies and performed by a variety of actors. A second version of the dataset includes four additional actions (drive car, eat, fight, run) and an increased number of samples for each class [83]. One training set is automatically annotated using scripts of the movies, another is manually labeled. There is a huge variety of performance of the actions, both spatially and temporally. Occlusions, camera movements and dynamic backgrounds make this dataset challenging. Most of the samples are at the scale of the upper-body but some show the entire body or a close-up of the face.

1.4.6. Other datasets

Ke et al. introduced the crowded videos dataset in [64]. Datasets containing still images figure skating, baseball and basketball are presented in [158]. İkizler et al. [54] presented a set of still images collected from the web.

2. Image representation

In this section, we discuss the features that are extracted from the image sequences. Ideally, these should generalize over small variations in person appearance, background, viewpoint and action execution. At the same time, the representations must be sufficiently rich to allow for robust classification of the action (see Section 3). The temporal aspect is important in action performance. Some of the image representations explicitly take into account the temporal dimension, others extract image features for each frame in the sequence individually. In this case, the temporal variations need to be dealt with in the classification step.

We divide image representations into two categories: global representations and local representations. The former encodes the visual observation as a whole. Global representations are obtained in a top-down fashion: a person is localized first in the image using background subtraction or tracking. Then, the region of interest is encoded as a whole, which results in the image descriptor. The representations are powerful since they encode much of the information. However, they rely on accurate localization, background subtraction or tracking. Also, they are more sensitive to viewpoint, noise and occlusions. When the domain allows for good control of these factors, global representations usually perform well.

Local representations describe the observation as a collection of independent patches. The calculation of local representations proceeds in a bottom-up fashion: spatio-temporal interest points are detected first, and local patches are calculated around these points. Finally, the patches are combined into a final representation. After initial success of bag-of-feature approaches, there is currently more focus on correlations between patches. Local representations are less sensitive to noise and partial occlusion, and do not strictly require background subtraction or tracking. However, as they depend on the extraction of a sufficient amount of relevant interest points, pre-processing is sometimes needed, for example to compensate for camera movements.

We discuss global and local image representations in Sections 2.1 and 2.2, respectively. A small number of works report the use of very specific features. We discuss these briefly in Section 2.3.

2.1. Global representations

Global representations encode the region of interest (ROI) of a person as a whole. The ROI is usually obtained through background subtraction or tracking. Common global representations are derived from silhouettes, edges or optical flow. They are sensitive to noise, partial occlusions and variations in viewpoint. To partly overcome these issues, grid-based approaches spatially divide the observation into cells, each of which encodes part of the observation locally (see Section 2.1.1). Multiple images over time can be stacked, to form a three-dimensional space–time volume, where time is the third dimension. Such volumes can be used for action recognition, and we discuss work in this area in Section 2.1.2.

The silhouette of a person in the image can be obtained by using background subtraction. In general, silhouettes contain some noise due to imperfect extraction. Also, they are somewhat sensitive to different viewpoints, and implicitly encode the anthropometry of the person. Still, they encode a great deal of information. When the silhouette is obtained, there are many different ways to encode either the silhouette area or the contour.

One of the earliest uses of silhouettes is by Bobick and Davis [11]. They extract silhouettes from a single view and aggregate differences between subsequent frames of an action sequence. This results in a binary motion energy image (MEI) which indicates where motion occurs. Also, a motion history image (MHI) is constructed where pixel intensities are a recency function of the silhouette motion. Two templates are compared using Hu moments. Wang et al. [162] apply a $\Re$ transform to extracted silhouettes. This results in a translation and scale invariant representation. Souvenir and Babbs [137] calculate a $\Re$ transform surface where the third dimension is time. Contours are used in [16], where the star skeleton describes the angles between a reference line, and the lines from the center to the gross extremities (head, feet, hands) of the contour. Wang and Suter [154] use both silhouette and contour descriptors. Given a sequence of frames, an average silhouette is formed by calculating the mean intensity over all centered frames. Similarly, the mean shape is formed from the centered contours of all frames. Weinland et al. [164] match two silhouettes using Euclidean distance. In later work [163], silhouette templates are matched against edges using Chamfer distance, thus eliminating the need for background subtraction.

When multiple cameras are employed, silhouettes can be obtained from each. Huang and Xu [52] use two orthogonally placed cameras at approximately similar height and distance to the person. Silhouettes from both cameras are aligned at the medial axis, and an envelope shape is calculated. Cherla et al. [17] also use orthogonally placed cameras and combine features of both. Such representations are somewhat view-invariant, but fail when the arms cannot be distinguished from the body. Weinland et al. [166] combine silhouettes from multiple cameras into a 3D voxel model. Such a representation is informative but requires accurate camera calibration. They use motion history volumes (see Fig. 2b), which is an extension of the MHI [11] to 3D. View-invariant matching is performed by aligning the volumes using Fourier transforms on the cylindrical coordinate system around the medial axis.

Instead of (silhouette) shape, motion information can be used. The observation within the ROI can be described with optical flow, the pixel-wise oriented difference between subsequent frames. Flow can be used when background subtraction cannot be performed. However, dynamic backgrounds can introduce noise in the motion descriptor. Also, camera movement results in observed motion, which can be compensated for by tracking the person. Efros et al. [27] calculate optical flow in person-centered images. They use sports footage, where persons in the image are very small. The result is blurred as optical flow can result in noisy displacement vectors. To make sure that oppositely directed vectors do not even out, the horizontal and vertical components are divided into positively and negatively directed, yielding 4 distinct channels. Ahad et al. [3] use these four flow channels to solve the issue of self-occlusion in a MHI approach. Ali and Shah [5] derive a number of kinematic features from the optical flow. These include divergence, vorticity, symmetry and gradient tensor features. Principal component analysis (PCA) is applied to determine dominant kinematic modes.

2.1.1. Global grid-based representations

By dividing the ROI into a fixed spatial or temporal grid, small variations due to noise, partial occlusions and changes in viewpoint can be partly overcome. Each cell in the grid describes the image observation locally, and the matching function is changed accordingly from global to local. These grid-based representations resemble local representations (see Section 2.2), but require a global representation of the ROI.

Kellokumpu et al. [66] calculate local binary patterns along the temporal dimension and store a histogram of non-background responses in a spatial grid. Thurau and Hlaváč [141] use histograms of oriented gradients (HOG, [23]) and focus on foreground edges by applying non-negative matrix factorization. Lu and Little [80] apply PCA after calculating the HOG descriptor, which greatly reduces the dimensionality. İkizler et al. [54] first extract human poses using [113]. Within the obtained outline, oriented rectangles are detected and stored in a circular histogram. Ragheb et al. [112] transform, for each spatial location, the binary silhouette response over time into the frequency domain. Each cell in the spatial grid contains the mean frequency response of the spatial locations it contains.

Optical flow in a grid-based representation is used by Danafar and Gheissari [24]. They adapt the work of Efros et al. [27] by dividing the ROI into horizontal slices that approximately contain head, body and legs. Zhang et al. [179] use an adaptation of the shape context, where each log-polar bin corresponds to a histogram of motion word frequencies. Combinations of flow and shape descriptors are also common, and overcome the limitations of a single representation. Tran et al. [142] use rectangular grids of silhouettes and flow. Within each cell, a circular grid is used to accumulate the responses. İkizler et al. [53] combine the work of Efros et al.

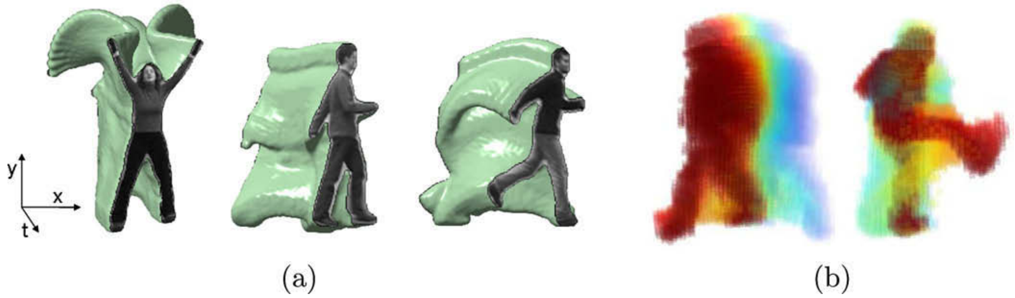

Fig. 2. (a) Space–time volume of stacked silhouettes (reprinted from [45], © IEEE, 2007) (b) Motion history volumes (reprinted from [166], © Elsevier, 2006). Even though the representations appear similar, (a) is viewed from a single camera, whereas (b) shows a recency function over reconstructed 3D voxel models.

[27] with histograms of oriented line segments. Flow, in combination with local binary patterns is used in [172].

2.1.2. Space–time volumes

A 3D spatio-temporal volume (STV) is formed by stacking frames over a given sequence. Accurate localization, alignment and possibly background subtraction are required.

Blank et al. [9,45] first stack silhouettes over a given sequence to form an STV (see Fig. 2a). Then the solution of the Poisson equation is used to derive local space–time saliency and orientation features. Global features for a given temporal range are obtained by calculating weighted moments over these local features. To deal with performances of different temporal durations, Achard et al. [1] use a set of space–time volumes for each sequence, each of which covers only a part of the temporal dimension.

Several works sample the STV surface and extract local descriptors. While this approach shares many similarities with local approaches, the STV is a global representation. Batra et al. [6] stack silhouettes, and sample the volume with small 3D binary space–time patches. Yilmaz and Shah [175] use differential geometric properties on the STV surface, such as maxima and minima in the space–time domain. An action sketch is the set of these local descriptors. The method is sensitive to noise on the surface. The idea is extended by Yan et al. [171] by first constructing 3D exemplars from multiple views, for each frame in a training sequence. Then, for each view, an action sketch is calculated from the view-based STV and projected onto the constructed 3D exemplars. The action sketch descriptors encode both shape and motion, and can be matched with observations from arbitrary viewpoints. Grundmann et al. [46] extend the shape context to 3D and apply it to STVs. The sampling of interest points is adapted to give more importance to moving regions.

Jiang and Martin [60] use 3D shape flows over time, calculated at edge points. The matching can deal with cluttered backgrounds. Ke et al. [63] construct an STV of flow and sample the horizontal and vertical components in space–time using a 3D variant of the rectangle features of [149]. Ogata et al. [100] extend this work with [27]. A combination of STVs of silhouettes and flow is used by Ke et al. [65]. No background subtraction is needed, as 3D super-pixels are obtained from segmenting the STV. Action classification is cast as 3D object matching, where the distance to the segment boundary is used as a similarity measure. The work is extended in [64] to allow for the matching of parts, thus enabling recognition of actions under partial occlusion.

2.2. Local representations

Local representations describe the observation as a collection of local descriptors or patches. Accurate localization and background subtraction are not required and local representations are somewhat invariant to changes in viewpoint, person appearance and partial occlusions. Patches are sampled either densely or at space–time interest points. The latter are locations that correspond to interesting motions and we discuss these in Section 2.2.1. Local descriptors describe small windows (2D) in an image or cuboids (3D) in a video volume, and are discussed in Section 2.2.2. Similar to global representations, observations can be grouped locally within a grid, see Section 2.2.3. By exploiting correlations in space and time between the patches, actions can be modeled more effectively since only the meaningful patches are retained. We discuss these correlations in Section 2.2.4.

2.2.1. Space–time interest point detectors

Space–time interest points are the locations in space and time where sudden changes of movement occur in the video. It is assumed that these locations are most informative for the recognition of human action. Usually, points that undergo a translational motion in time will not result in the generation of space–time interest points.

Laptev and Lindeberg [72] extended the Harris corner detector [49] to 3D. Space–time interest points are those points where the local neighborhood has a significant variation in both the spatial and the temporal domain. The scale of the neighborhood is automatically selected for space and time individually. The work is extended to compensate for relative camera motions in [71]. Oikonomopoulos et al. [102] extended the work on 2D salient point detection by Kadir and Brady [62] to 3D. The entropy within each cuboid is calculated, and the centers of those with local maximum energy are selected as salient points. The scale of each salient point is determined by maximizing the entropy values.

One drawback of these methods is the relatively small number of stable interest points. This issue is addressed by Dollár et al. [25], who apply Gabor filtering on the spatial and temporal dimensions individually. The number of interest points is adjusted by changing the spatial and temporal size of the neighborhood in which local minima are selected. Chomat et al. [19] use the responses after applying spatio-temporal receptive fields. In a similar fashion, Rapantzikos et al. [118] apply discrete wavelet transforms in each of the three directions of a video volume. Responses from low-pass and high-pass filtering for each dimension are used to select salient points in space and time. In addition to intensity and motion cues, Rapantzikos et al. [117] also incorporate color. They compute saliency as the solution of an energy minimization process which involves proximity, scale and feature similarity terms.

Willems et al. [167] identify saliency as the determinant of a 3D Hessian matrix, which can be calculated efficiently due to the use of integral videos. Another attempt to decrease the computational complexity is presented by Oshin et al. [103], who train randomized ferns to approximate the behavior of interest point detectors. In a comparison on the task of human action recognition, Wang et al. [151] found that dense sampling outperformed the interest point detectors of Dollár et al. [25], Laptev and Lindeberg [72] and Willems et al. [167].

Instead of detecting interest points over the entire volume, Wong and Cipolla [168] first detect subspaces of correlated movement. These subspaces correspond to large movements such as an arm wave. Within these spaces, a sparse set of interest points is detected. In a similar approach, Bregonzio et al. [13] first calculate the difference between subsequent frames to estimate the focus of attention. Next, Gabor filtering is used to detect salient points within these regions.

2.2.2. Local descriptors

Local descriptors summarize an image or video patch in a representation that is ideally invariant to background clutter, appearance and occlusions, and possibly to rotation and scale. The spatial and temporal size of a patch is usually determined by the scale of the interest point. Fig. 3 shows cuboids at detected interest points. Schüldt et al. [125] calculate patches of normalized derivatives in space and time. Niebles et al. [95] take the same approach, but apply smoothing before reducing the dimensionality using PCA. Dollár et al. [25] experiment with both image gradients and optical flow.

Patches can also be described by local grid-based descriptors. These summarize the local observation within grid cells, thus ignoring small spatial and temporal variations. SURF features [7] are extended to 3D by Willems et al. [167]. These eSURF features contain in each cell the sums of Haar-wavelets. Laptev et al. [73] use local HOG and HOF (histogram of oriented flow) descriptors. The extension of HOG to 3D is presented by Kläser et al. [68]. 3D gradients are binned into regular polyhedrons. They extend the idea of integral images into 3D which allows rapid dense sampling

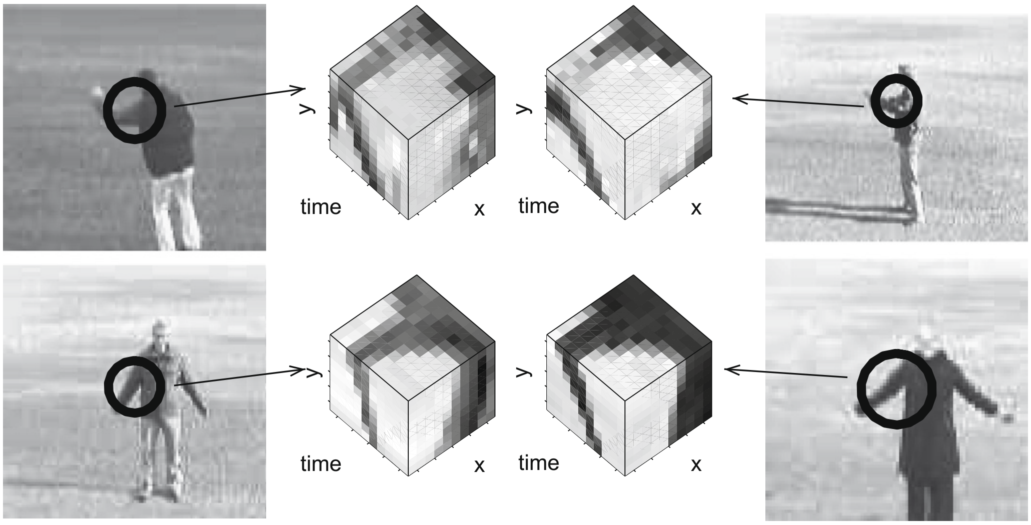

Fig. 3. Extraction of space–time cuboids at interest points from similar actions performed by different persons (reprinted from [71], © Elsevier, 2007).

of the cuboid over multiple scales and locations in both space and time. In related work by Scovanner et al. [126], the SIFT descriptor [79] is extended to 3D. Wang et al. [151] compared local descriptors and found that, in general, a combination of image gradient and flow information resulted in the best performance.

Several approaches combine interest point detection and the calculation of local descriptors in a feed-forward framework. For example, Jhuang et al. [58] use several stages to ensure invariance to a number of factors. Their approach is motivated by the human visual system. At the lowest level, Gabor filters are applied to dense flow vectors, followed by a local max operation. Then the responses are converted to a higher level using stored prototypes and a global max operation is applied. A second matching stage with prototypes results in the final representation. The work in [96] is similar in concept, but uses different window settings. Schindler and Van Gool [124] extend the work by Jhuang et al. [58] by combining both shape and flow responses. Escobar et al. [31] use motion-sensitive responses and also consider interactions between cells, which allows them to model more complex properties such as motion contrasts.

Comparing sets of local descriptors is not straightforward due to the possibly different number and the usually high dimensionality of the descriptors. Therefore, often a codebook is generated by clustering patches and selecting either cluster centers or the closest patches as codewords. A local descriptor is described as a codeword contribution. A frame or sequence can be represented as a bag-of-words, a histogram of codeword frequencies (e.g. [95,125]).

2.2.3. Local grid-based representations

Similar to holistic approaches, described in Section 2.1.1, grids can be used to bin the patches spatially or temporally. Compared to the bag-of-words approach, using a grid ensures that spatial information is maintained to some degree.

In the spatial domain, İkizler and Duygulu [56] sample oriented rectangular patches, which they bin into a grid. Each cell has an associated histogram that represents the distribution of rectangle orientations. Zhao and Elgammal [180] bin local descriptors around interest points in a histogram with different levels of granularity. Patches are weighted according to their temporal distance to the current frame.

Nowozin et al. [98] use a temporal instead of a spatial grid. The cells overlap, which allows them to overcome small variations in performance. Observations are described as PCA-reduced vectors around extracted interest points, mapped onto codebook indices.

Laptev and Pérez [74] bin histograms of oriented gradients and flow, extracted at interest points, into a spatio-temporal grid. This grid spans the volume that is determined based on the position and size of a detected head. The distribution of these histograms is determined for every spatio-temporal cell in the grid. Three different block types are used to form the new feature set. These types correspond to a single cell, a concatenation of two temporally neighboring cells and a concatenation of spatially neighboring cells. A subset of all possible blocks within the grid is selected using AdaBoost. A larger number of grid types, with different spatial and temporal divisions and overlap settings, is evaluated in [73]. Flow descriptors from [27] are used by Fathi and Mori [35], who select a discriminative set of low-level flow features within space–time cells which form an overlapping grid. In a subsequent step, a set of these mid-level features is selected using the AdaBoost algorithm. In the work by Bregonzio et al. [13], no local image descriptor are calculated. Rather, they look at the number of interest points within cells of a spatio-temporal grid with different scales. This approach is computationally efficient but depends on the number and relevancy of the interest points.

2.2.4. Correlations between local descriptors

Grid-based representations model spatial and temporal relations between local descriptors to some extent. However, they are often redundant and contain uninformative features. In this section, we describe approaches that exploit correlations between local descriptors for selection or the construction of higher-level descriptors.

Scovanner et al. [126] construct a word co-occurrence matrix, and iteratively merge words with similar co-occurrences until the difference between all pairs of words is above a specified threshold. This leads to a reduced codebook size and similar actions are likely to generate more similar distributions of codewords. Similar in concept is the work by Liu et al. [76], who use a combination of the space–time features and spin images, which globally describe an STV. A co-occurrence matrix of the features and the action videos is constructed. The matrix is decomposed into eigenvectors and subsequently projected onto a lower-dimensional space. This embedding can be seen as feature-level fusion. Instead of determining pairs of correlated codewords, Patron-Perez and Reid [106] approximate the full joint distribution of features using first-order dependencies. Features are binary variables that indicate the presence of a codeword. A maximum spanning tree is formed by analyzing a graph between all pairs of features. The work by Kim et al. [67] is different in the sense that the correlation between two videos is measured. Canonical correlation analysis is extended to handle image sequences. The approach implicitly deals with affine variations. Discriminative features are subsequently selected using AdaBoost.

In contrast to the above approaches where spatial information is ignored, Savarese et al. [123] introduce correlatons that describe co-occurrences of codewords within spatio-temporal neighborhoods. The codebook size strongly influences the classification performance. Too few entries do not allow for good discrimination, while too great a codebook size is likely to introduce noise due to sparsity of the histograms. Liu and Shah [78] solve this issue and determine the optimal size of the codebook using maximization of mutual information. This technique merges two codebook entries if they have comparable distributions. In addition, they use spatio-temporal pyramid matching to exploit temporal information. Yao and Zhu [173] introduce an active basis of shape and flow patches, where locations in space and time are allowed to vary slightly.

Correlations between descriptors can also be obtained by tracking features. Sun et al. [140] calculate SIFT descriptors around interest points in each frame and use Markov chaining to determine tracks of these features. Similar work by Messing et al. [87] extracts trajectories using the KLT tracker. In both cases, tracks are summarized in a log-polar histogram of track velocities. Oikonomopoulos et al. [101] fit B-splines to the STV boundary that is formed by a coherent region of saliency responses. Song et al. [136] track points between frames. They fit a triangulated graph to these points to detect and recognize human actions. In Fanti et al. [32], additional local appearance cues are used. Global variables are introduced for scale, viewpoint and translation. These methods assume static backgrounds and motion due to objects in the background generates feature tracks that do not belong to the person.

This limitation is partly addressed by Niebles and Fei-Fei [94], who model the frame as a mixture of constellations. Each constellation models the spatial arrangement of codewords instead of tracked features. Filipovych and Ribeiro [37] include both pose constellations and dynamics constellations. Star graphs of static and dynamic features are combined into a tree by conditioning on the landmark vertices of the individual graphs. These models are trained without supervision. Related work introduces hidden variables that correspond to action categories. Probabilistic latent semantic analysis (pLSA) is a generative model used by Niebles et al. [95]. In an unsupervised way, the mapping from latent action labels to distribution of codewords is learned. Wong et al. [169] extend pLSA by including the location of a person's centroid. Both works require that the number of action labels is determined empirically. Instead, Wang et al. [160] take a supervised approach and use a semi-latent Dirichlet allocation (S-LDA) model. In Wang and Mori [159], an adapted hidden conditional random field (HCRF) model is used to learn constellations of codewords discriminatively. A more efficient learning algorithm is presented in [161].

A number of works take the approach of extracting and mining large number of features. Mikolajczyk and Uemura [88] extract for each frame local shape and motion features. For each feature, the relative location and orientation to a person's center of mass are stored together with the annotated action label. These features are clustered and represented in vocabulary trees. By matching features extracted from an observed frame, votes are cast over a persons' location, orientation and action label. In Uemura et al. [146], global motion patterns are detected and compensated for, in order to recognize action from moving cameras. In related work, Gilbert et al. [44] find spatio-temporal corners and determine the relative spatial arrangement of all other corners in the frame. This results in an extremely large number of features. Data mining techniques are further used to discriminatively select those feature combinations that are informative of a class. Liu et al. [77] select discriminative features by applying the PageRank algorithm on the feature co-occurrence graph.

There are relatively few works that address the effect of viewpoint on the recognition of human actions. Farhadi and Tabrizi [33] explicitly address the correlations between actions observed from different views. They use a split-based representation to describe clusters of codewords in each view. The transfer of these splits between views is learned from multi-view action sequences. Farhadi et al. [34] model the view as a latent parameter, and learn features that can discriminate between views and actions.

2.3. Application-specific representations

In contrast to the more general image representations that have been discussed in the previous sections, a number of works use representations that are directly motivated by the domain of human action recognition.

Joint locations or joint angles are rich representations, but it is challenging to derive them from video (see [38,109]). In 3D, the representations are completely view-invariant, whereas for 2D, there have been several approaches proposed to address the issue of matching 2D joint trajectories to action labels (e.g. [5,104,116,131,132,174]). Since we focus on the recognition of human actions from image and video, we do not discuss these works here.

Smith et al. [135] use a number of specifically selected features. Some of these are low-level and deal with color and movement. Others are higher-level and are obtained from head and hand regions. A boosting scheme is used that takes into account the history of the action performance. The work by Vitaladevuni et al. [150] is inspired by the observation that human actions differ in accelerating and decelerating force. They identify reach, yank and throw types. Temporal segmentation into atomic movements described with movement type, spatial location and direction of movement is performed first.

3. Action classification

When an image representation is available for an observed frame or sequence, human action recognition becomes a classification problem. An action label or distribution over labels is given for each frame or sequence. Section 3.1 discusses approaches that classify image representations into actions without explicitly modeling variations in time. Temporal state-space approaches do model such variations of an action and are discussed in Section 3.2. In Section 3.3, we describe general approaches to detect human action in video without modeling the action.

3.1. Direct classification

The approaches that we describe in this section do not pay special attention to the temporal domain. They summarize all frames of an observed sequence into a single representation or perform action recognition for each frame individually. While both these approaches can deal with variations in executing and recording rate, the temporal order is neglected. In Section 3.1.2, we discuss nearest neighbor classification where an observed sequence is compared to labeled sequences or action class prototypes. A second class of approach is that of the discriminative classifiers. These learn a function that discriminates between two or more classes by directly operating on the image representation. Dimensionality reduction is a common step before the actual classification and is discussed first.

3.1.1. Dimensionality reduction

In many cases, image representations are high-dimensional. This makes matching computationally more expensive. Also, the representation might contain noisy features. It is expected that a more compact, robust feature representation is obtained by embedding the space of image representations onto a lower dimensional space. This embedding can be learned from training data.

PCA is a common linear dimensionality reduction method that has been used by Masoud and Papanikolopoulos [84] and Rosales [121]. Often, the mapping between full and lower dimensional image representation is better described as a non-linear function. Chin et al. [18] learn manifolds using local linear embedding (LLE). They experiment with different projection functions. Wang and Suter [157] use locality preserving projections (LPP), Isomap is used by Blackburn and Ribeiro [8].

The above dimensionality reduction methods learn the embedding in an unsupervised manner, which does not guarantee good discrimination between classes. Poppe and Poel [110] address this issue and learn discriminative feature transforms between pairs of classes. Jia and Yeung [59] use an embedding that is discriminative both in a spatial and temporal sense. They propose local spatio-temporal discriminant embedding (LSTDE), which maps silhouettes of the same class close in the manifold and model temporal relations in subspaces of the manifold.

3.1.2. Nearest neighbor classification

k-Nearest neighbor (NN) classifiers use the distance between the image representation of an observed sequence and those in a training set. The most common label among the *k* closest training sequences is chosen as the classification. For a large training set, such comparisons can be computationally expensive. Alternatively, for each class, an action prototype can be calculated by taking the mean of all corresponding sequences. The ability to cope with variations in spatial and temporal performance, viewpoint and image appearance depends on the training set, the type of image representation and the distance metric.

NN classification can be either performed at the frame level, or for whole sequences. In the latter case, issues with different frame lengths need to be resolved, for example by using majority voting over all frames in a sequence. 1-NN with Euclidean distance are used by Blank et al. [9] for global features and Batra et al. [6] for histograms of codewords. Euclidean distance might not be the most suitable choice given the type of image representation. Bobick and Davis [11] use Hu moments of different orders of magnitude. Mahalanobis distance is used to take into account the variance of each dimension. Rodriguez et al. [120] describe a method to generate spatio-temporal templates that effectively capture the intra-class variance into a single prototype.

Several authors have used NN classification in combination with dimensionality reduction. Wang and Suter [155] either use the minimum mean frame-wise distance in an embedded space, or a frame-order preserving variant. Turaga et al. [145] focus on parametric and non-parametric manifold density functions and describe distance functions for Grassmann and Stiefel manifold embeddings. Tran et al. [142] and Poppe and Poel [110] use a learned discriminative distance metric in the NN classification.

It has been observed that many actions can be represented by key poses or prototypes. Sullivan and Carlsson [138] recognize forehand and backhand tennis strokes by matching edge representations to labeled key poses. Wang et al. [158] also use edge representations but learn action clusters in an unsupervised fashion. They manually provide action class labels after the clustering. Weinland et al. [165] learn a set of action key poses as 3D voxel representations. These methods use only a single frame for action classification. As many poses are only weakly informative for the action class, considering a sequence of poses over time is likely to reduce ambiguities. Weinland and Boyer [163] use the minimum distance of each key pose to the frames in the sequences. The set of key poses is discriminatively selected. Lin et al. [75] store prototypes in a tree to allow for efficient matching.

3.1.3. Discriminative classifiers

Discriminative classifiers focus on separating two or more classes, rather than modeling them. Support vector machines (SVM) learn a hyperplane in feature space that is described by a weighted combination of support vectors. SVMs have been used in combination with local representations of fixed lengths, such as histograms of codewords in [58,71,125]. Relevance vector machines (RVM) can be regarded as the probabilistic variant of the SVM. Training an RVM usually results in a sparser set of support vectors. They have been used for action recognition by Oikonomopoulos et al. [102].

In a boosting framework, a final strong classifier is formed by a set of weak classifiers, each of which usually uses only a single dimension of the image representation. Boosting is used in many works, either as a discriminative feature selection step or as the actual classifier. AdaBoost [39] has been used by [35,74,100]. LPBoost yields sparser coefficients and is reported to converge faster, and is used in [98]. Smith et al. [135] introduce a variant that uses history information in the boosting scheme.

3.2. Temporal state-space models

State-space models consist of states connected by edges. These edges model probabilities between states, and between states and observations. In the models that we discuss in this section, each state summarizes the action performance at a certain moment in time. An observation corresponds to the image representation at a given time. Temporal state-space models are either generative or discriminative. While they share many characteristics, they are conceptually different. Generative models learn a joint distribution over both observations and action labels. They thus learn to model a certain action class, with all its variations. In contrast, discriminative models learn probabilities of the action classes conditioned on the observations. They do not model a class but rather focus on differences between classes. We discuss generative and discriminative models in Sections 3.2.2 and 3.2.3, respectively. Dynamic Time Warping (DTW) can be regarded as a generative model, but it is used between pairs of sequences. Due to this rather different use, we discuss DTW separately in Section 3.2.1.

3.2.1. Dynamic time warping

Dynamic time warping is a distance measure between two sequences, possibly with different lengths. It simultaneously takes into account a pair-wise distance between corresponding frames and the sequence alignment cost. For a low alignment cost, two sequences need to be segmented similarly in time and be performed at similar rates. Dynamic programming is used to calculate the optimal alignment. Veeraraghavan et al. [148] use DTW for sequences of normalized shape features. As these lie on a spherical manifold, the distance function between shapes is adapted. In [147], they also address the alignment by considering the space of temporal warping functions for a given activity. Yao et al. [173] introduce dynamic space–time warping where, in addition to the temporal dimension, sequences are also aligned on image position and scale. A related distance is longest common subsequence (LCS). It only takes into account similar elements of both sequences and results in an increased distance when more inserts or deletions are necessary. LCS has been used in [50,172].

3.2.2. Generative models

Hidden Markov models (HMM) use hidden states that correspond to different phases in the performance of an action. They model state transition probabilities and observation probabilities. To keep the modeling of the joint distribution over representation and labels tractable, two independence assumptions are introduced. First, state transitions are conditioned only on the previous state, not on the state history. This is the Markov assumption. Second, observations are conditioned only on the current state, so subsequent observations are considered independent.

HMMs have been used in a large number of works. Yamato et al. [170] cluster grid-based silhouette mesh features to form a compact codebook of observations. They train HMMs for the recognition of different tennis strokes. Training of an HMM can be done efficiently using the Baum–Welch algorithm. The Viterbi algorithm is used to determine the probability of observing a given sequence. When using a single HMM per action, action recognition becomes finding the action HMM that could generate the observed sequence with the highest probability.

Feng and Perona [36] use a static HMM where keyposes correspond to states. They effectively train the dynamics at the cost of reduced flexibility due to a simpler observation model. Weinland et al. [164] construct a codebook by discriminatively selecting templates. In the HMM, they condition the observation on the viewpoint. Related work by Lv and Nevatia [82] uses an Action Net, which is constructed by considering key poses and viewpoints. Transitions between views and poses are encoded explicitly. Ahmad and Lee [4] take into account multiple viewpoints and use a multi-dimensional HMM to deal with the different observations. Instead of modeling viewpoint, Lu and Little [80] use a hybrid HMM where one process denotes the closest shape-motion template while the other models position, velocity and scale of the person in the image. Ramanan and Forsyth [115] track persons in 2D by learning the appearance of the body-parts. In [114], these 2D tracks are subsequently lifted to 3D using stored snippets of annotated pose and motion. An HMM is used to infer the action from these labeled codeword motions.

Instead of modeling the human body as a single observation, one HMM can be used for each every body-part. This makes training easier, as the combinatorial complexity is reduced to learning dynamical models for each limb individually. In addition, composite movements that are not in the training set can be recognized. İkizler and Forsyth [57] use the 3D body-part trajectories that are obtained using [114]. They construct HMMs for the legs and arms individually, where 3D trajectories are the observations. For each limb, states of different action models with similar emission probabilities are linked, which allows for automatic segmentation of actions. A similar approach has been taken by Chakraborty et al. [15], where arms, legs and head are found with a set of view-dependent detectors. Lv and Nevatia [81] also use 3D joint locations but construct a large number of action HMMs, each of which uses a subset of all joints. This results in a large number of weak classifiers. They use AdaBoost to form the final classifier.

Several works aim at improving pose recovery by modeling the dynamics for each class of movement. These approaches can also be used for action recognition by selecting the action class whose corresponding model has the highest probability of generating the observed sequence. Peursum et al. [107] use a factored-state hierarchical HMM (FS-HHMM) to jointly model image observations and body dynamics per action class. Caillette et al. [14] uses a variable length Markov model (VLMM) to model observations and 3D poses for each action. Natarajan and Nevatia [92] introduce a hierarchical variable transition HMM (HVT–HMM) which consists of three layers that model composite actions, primitive actions and poses. Due to the variable window, actions can be recognized with low latency.

Grammars are generative models that specify explicitly in which order parts of an action can be observed. Ogale et al. [99] construct a probabilistic context-free grammar where probabilities of pose pairs are learned from training data, while small viewpoint changes are allowed. Turaga et al. [144] model an action as a cascade of linear time invariant (LTI) dynamical models. In an unsupervised way, they simultaneously learn the dynamical models and temporally segment a sequence. Similar models are grouped into action prototypes. A cascade structure is formed by learning *n*-grams over the sequence of action prototypes. This cascade can be regarded as a grammar that describes the production rules for each action in terms of a sequence of action prototypes.

3.2.3. Discriminative models

The independence assumptions in HMMs assume that observations in time are independent, which is often not the case. Discriminative models overcome this issue by modeling a conditional distribution over action labels given the observations. These models can take into account multiple observations on different timescales. They can be trained to discriminate between action classes rather than learning to model each class individually, as in generative models. Discriminative models are suitable for classification of related actions that could easily be confused using a generative approach. In general, discriminative graphical models require many training sequences to robustly determine all parameters.

Conditional random fields (CRF) are discriminative models that can use multiple overlapping features. Sminchisescu et al. [134] use a linear chain CRF, where the state dependency is first-order. They compare CRFs with HMMs and maximum entropy Markov models (MEMM). The latter are related to the CRF, but are directed models instead. They suffer from the label bias problem, where states with few outgoing transitions are favored. A more detailed comparison between CRFs and MEMMs is given in [70]. Sminchisescu et al. show that CRFs outperform both MEMMs and HMMs when using larger windows, which take into account more of the observation history. These results are partly supported by Mendoza and Pérez de la Blanca [86], who obtain better results for CRFs compared to HMMs using shape features, especially for related actions (e.g. walking and jogging). Interestingly, when using motion features, HMMs outperformed CRFs.

Variants of CRFs have also been proposed. Wang and Suter [156] use a factorial CRF (FCRF), a generalization of the CRF. Structure and parameters are repeated over a sequence of state vectors, which can be regarded as a distributed state representation. This allows for the modeling of complex interactions between labels and long-range dependencies, while inference is approximate instead of exact as in CRFs. Zhang and Gong [178] use a hidden CRF (HCRF, [111]) to label sequences as a whole. They introduce a HMM pathing stage, which ensures that learning the HCRF parameters is globally optimal. Natarajan and Nevatia [93] use a two-layer model where the top level encodes action and viewpoint. On the lower level, CRFs are used to encode the action and viewpoint-specific pose observation. Ning et al. [97] combine a discriminative pose recovery approach with a CRF for action recognition. The parameters of both layers are jointly optimized. No image-to-pose data is required during training, but has been shown to improve performance. Shi et al. [133] use a semi-Markov model (SMM), which is suitable for both action segmentation and action recognition.

3.3. Action detection

Action detection approaches do not explicitly model the image representation of a person in the image, nor do they model action dynamics. Rather, they correlate an observed sequence to labeled video sequences. Such work is aimed at the detection of actions, rather than at their recognition. However, these works share many similarities to those previously discussed and we will describe them briefly in this section. The detection of cyclic actions is discussed in Section 3.3.1.

Zelnik-Manor and Irani [177] describe video segments as bag-of-words encoded over different temporal scales. Each word is the gradient orientation of a local patch. Patches with low temporal variance are ignored, which focuses the representation on moving areas. This restricts the approach to detection against non-moving backgrounds. Ning et al. [96] use Gabor responses instead of gradient orientations. In both works, a histogram distance measure is used.

Shechtman and Irani [130] consider the spatial dimension by correlating space–time patches over different locations in space and time. They use patches that locally describe motion. To avoid calculating the optical flow, a rank-based constraint is used directly on the intensity information of the cuboids. Matikainen et al. [85] approximate this approach but use motion words and a look-up table to allow for faster correlation. More recently, Shechtman and Irani [129], propose a self-similarity descriptor that correlates local patches. The descriptor is invariant to color, texture and can deal with small spatial variations. A query template is described by an ensemble of all descriptors. Seo et al. [128] use space–time local steering kernels, which can be regarded as a generalization of [129]. By applying PCA on a collection of kernels, they obtain the most salient features.

The above methods require that a window is sled through time and space, which makes them computationally expensive. This issue is addressed by Hu et al. [51], who describe a sequence as a collection of windows of different temporal scales and positions and use multiple-instance learning to learn the binary action classifier. Yuan et al. [176] detect space–time interest points and classify whether each is part of the query action. An efficient branch-and-bound approach is taken to search for the subvolume that has the maximum of positively labeled points.

Junejo et al. [61] observe that the temporal self-similarity matrix of an action seen from different viewpoints is very similar (see Fig. 4). They describe a sequence as a histogram of local descriptors, calculated from the self-similarity matrix. Boiman and Irani [12] take a different approach by describing a sequence as an ensemble of local spatial or spatio-temporal patches. A similarity score is based on the composition of a query sequence from these patches. Similar sequences require less but larger patches.

3.3.1. Cyclic actions

Some works assume motion periodicity, which allows for temporal segmentation by analyzing the self-similarity matrix. Seitz and Dyer [127] introduce a periodicity detection algorithm that is able to cope with small variations in the temporal extent of a motion. They track markers and use an affine distance function. Cutler and Davis [21] perform a frequency transform on the self-similarity matrix of a tracked object. Peaks in the spectrum correspond to the frequency of the motion. The type of action is determined by analyzing the matrix structure. Polana and Nelson [108] also use Fourier transforms to find the periodicity and temporally segment the video. They match motion features to labeled 2D motion templates.

4. Discussion

In this section, we summarize the state of the art, point out limitations and identify promising directions for future research to address these limitations.

Global image representations have proven to yield good results, and they can usually be extracted with low cost. However, their applicability is limited to scenarios where ROIs can be determined reliably. Moreover, they cannot deal with occlusions. Local representations address these issues. Initial work used bag-of-feature representations but more recent work takes into account spatial and temporal correlations between patches. Still, the question how to deal with more severe occlusions has largely been ignored.

Most of the reported work is restricted to fixed and known viewpoints, which severely limits its applicability. The use of multiple view-dependent action models solves this issue but at the cost of increased training complexity. Recently, researchers have begun to address the recognition of actions from viewpoints for which there is no corresponding training data [33,34,61].

Regarding classification, we discussed direct classification and temporal state-space models. In the former, temporal variations are not explicitly modeled, which proved to be a reasonable approach in many cases. For more complex motions, it is questionable whether this approach is suitable. Generative state-space models such as HMMs can model temporal variations but have difficulties distinguishing between related actions (e.g. jogging and walking). In this respect, discriminative graphical approaches are more suitable. In future work, the flexibility of the classifier with respect to adding or removing action classes from the repertoire will play a more important role.

Many approaches assume that the video is readily segmented into sequences that contain one instance of a known set of action labels. Often, it is also assumed that the location and approximate scale of the person in the video is known or can easily be estimated. The action detection task is thus ignored, which limits the applicability to situations where segmentation in space and time is possible. While several works (e.g. [51,176]) have addressed this topic, it remains a challenge to perform action detection for online applications.

Another aspect of human action recognition is the current evaluation practice. Publicly available datasets (see Section 1.4) have shaped the domain by allowing for objective comparison between approaches on common training and test data. They also allow for

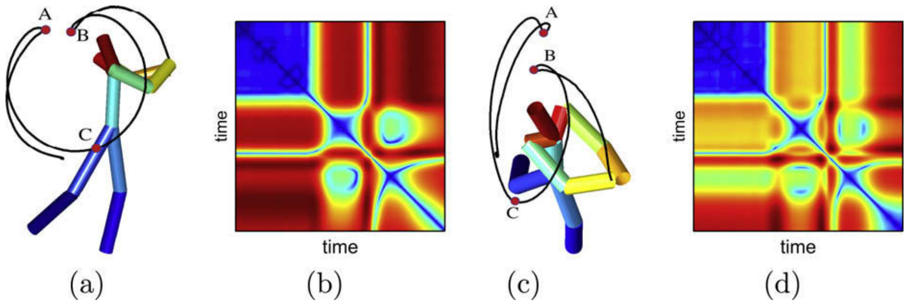

Fig. 4. Example of cross-correlation between viewpoints, (a and c) a golf swing seen from two different viewpoints, (b and d) the corresponding self-similarity matrices. Note the similarity in structure (reprinted from [61], © Springer-Verlag, 2008).

better understanding of methods since researchers are aware of the challenges of each set. However, algorithms may be biased to a particular dataset. This may lead to complex approaches that perform better on a specific dataset but may be less generally applicable.

Also, given the increasing level of sophistication of action recognition algorithms, larger and more complex datasets should direct research efforts to realistic settings. Initially, datasets were not focused on an application domain. However, action recognition in surveillance, human–computer interaction and video retrieval poses different challenges. Human–computer interaction applications require real-time processing, missed detections in surveillance are unacceptable and video retrieval applications often cannot benefit from a controlled setting and require a query interface (e.g. [139]). Currently, there is a shift towards a diversification in datasets. The HOHA dataset [73] targets action recognition in movies, whereas the UFC sport dataset [120] contains sport footage. Such a diversification is beneficial as it allows for realistic recording settings while focusing on relevant action classes. Moreover, the use of application-specific datasets allows for the use of evaluation metrics that go beyond precision and recall, such as speed of processing or detection accuracy. Still, the compilation or recording of datasets that contain sufficient variation in movements, recording settings and environmental settings remains challenging and should continue to be a topic of discussion.

Related is the issue of labeling data. For increasingly large and complex datasets, manual labeling will become prohibitive. Automatic labeling using video subtitles [48] and movie scripts [20,26,73] is possible in some domains, but still requires manual verification. When using an incremental approach to image harvesting such as in [55], the initial set will largely affect the final variety of action performances.

We discussed vision-based human action recognition in this survey but a multi-modal approach could improve recognition in some domains, for example in movie analysis. Also, context such as background, camera motion, interaction between persons and person identity provides informative cues [83].

Given the current state of the art and motivated by the broad range of applications that can benefit from robust human action recognition, it is expected that many of these challenges will be addressed in the near future. This would be a big step towards the fulfillment of the longstanding promise to achieve robust automatic recognition and interpretation of human action.

Acknowledgements

This work was supported by the European IST Programme Project FP6-033812 (Augmented Multi-party Interaction with Distant Access), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024. The author wishes to thank the reviewers for their valuable comments and the authors that contributed figures to this survey.

References

[1] Catherine Achard, Xingtai Qu, Arash Mokhber, Maurice Milgram, A novel approach for recognition of human actions with semi-global features, Machine Vision and Applications 19 (1) (2008) 27–34.
[2] Jake K. Aggarwal, Qin Cai, Human motion analysis: a review, Computer Vision and Image Understanding (CVIU) 73 (3) (1999) 428–440.
[3] Md. Atiqur Rahman Ahad, Takehito Ogata, Joo Kooi Tan, Hyoungseop Kim, Seiji Ishikawa, Motion recognition approach to solve overwriting in complex actions, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08), September 2008, Amsterdam, The Netherlands, 2008, pp. 1–6.
[4] Mohiuddin Ahmad, Seong-Whan Lee, Human action recognition using shape and CLG-motion flow from multi-view image sequences, Pattern Recognition 41 (7) (2008) 2237–2252.
[5] Saad Ali, Mubarak Shah, Human action recognition in videos using kinematic features and multiple instance learning, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), in press.
[6] Dhruv Batra, Tsuhan Chen, Rahul Sukthankar, Space–time shapelets for action recognition, in: Proceedings of the Workshop on Motion and Video Computing (WMVC'08), Copper Mountain, CO, January 2008, pp. 1–6.
[7] Herbert Bay, Andreas Ess, Tinne Tuytelaars, Luc J. Van Gool, SURF: Speeded up robust features, Computer Vision and Image Understanding (CVIU) 110 (3) (2008) 346–359.
[8] Jaron Blackburn, Eraldo Ribeiro, Human motion recognition using Isomap and dynamic time warping, in: Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07), Lecture Notes in Computer Science, Rio de Janeiro, Brazil, October 2007, pp. 285–298 (Number 4814).
[9] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, Ronen Basri, Actions as space–time shapes, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 2, Beijing, China, October 2005, pp. 1395–1402.
[10] Aaron F. Bobick, Movement, activity and action: the role of knowledge in the perception of motion, Philosophical Transactions of the Royal Society B: Biological Sciences 352 (1358) (1997) 1257–1265.
[11] Aaron F. Bobick, James W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 23 (3) (2001) 257–267.
[12] Oren Boiman, Michal Irani, Detecting irregularities in images and in video, International Journal of Computer Vision (IJCV) 74 (1) (2007) 17–31.
[13] Matteo Bregonzio, Shaogang Gong, Tao Xiang, Recognising action as clouds of space–time interest points, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.
[14] Fabrice Caillette, Aphrodite Galata, Toby Howard, Real-time 3-D human body tracking using learnt models of behaviour, Computer Vision and Image Understanding (CVIU) 109 (2) (2008) 112–125.
[15] Bhaskar Chakraborty, Ognjen Rudovic, Jordi Gonzàlez, View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08), September 2008, Amsterdam, The Netherlands, 2008, pp. 1–6.
[16] Hsuan-Sheng Chen, Hua-Tsung Chen, Yi-Wen Chen, Suh-Yin Lee, Human action recognition using star skeleton, in: Proceedings of the International Workshop on Video Surveillance and Sensor Networks (VSSN'06), Santa Barbara, CA, October 2006, pp. 171–178.
[17] Srikanth Cherla, Kaustubh Kulkarni, Amit Kale, Viswanathan Ramasubramanian, Towards fast, view-invariant human action recognition, in: Proceedings of the Workshop on Computer Vision and Pattern Recognition for Human Communicative Behavior Analysis (CVPR4HB'08), Anchorage, AK, June 2008, pp. 1–8.
[18] Tat-Jun Chin, Liang Wang, Konrad Schindler, David Suter, Extrapolating learned manifolds for human activity recognition, in: Proceedings of the International Conference on Image Processing (ICIP'07), vol. 1, San Antonio, TX, September 2007, pp. 381–384.
[19] Olivier Chomat, Jérôme Martin, James L. Crowley, A probabilistic sensor for the perception and recognition of activities, in: Proceedings of the European Conference on Computer Vision (ECCV'00), Lecture Notes in Computer Science, vol. 1, Dublin, Ireland, June 2000, pp. 487–503 (Number 1842).
[20] Timothee Cour, Chris Jordan, Eleni Miltsakaki, Ben Taskar, Movie/script: Alignment and parsing of video and text transcription, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 4, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 158–171 (Number 5305).
[21] Ross Cutler, Larry S. Davis, Robust real-time periodic motion detection, analysis, and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 22 (8) (2000) 781–796.
[22] Fabio Cuzzolin, Using bilinear models for view-invariant action and identity recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, New York, NY, June 2006, pp. 1701–1708.
[23] Navneet Dalal, Bill Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, San Diego, CA, June 2005, pp. 886–893.
[24] Somayeh Danafar, Niloofar Gheissari, Action recognition for surveillance applications using optic flow and SVM, in: Proceedings of the Asian Conference on Computer Vision (ACCV'07) – part 2, Lecture Notes in Computer Science, Tokyo, Japan, November 2007, pp. 457–466 (Number 4844).
[25] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, Serge Belongie, Behavior recognition via sparse spatio-temporal features, in: Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05), Beijing, China, October 2005, pp. 65–72.
[26] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, Jean Ponce, Automatic annotation of human actions in video, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.
[27] Alexei A. Efros, Alexander C. Berg, Greg Mori, Jitendra Malik, Recognizing action at a distance, in: Proceedings of the International Conference on Computer Vision (ICCV'03), vol. 2, Nice, France, October 2003, pp. 726–733.

[28] Ahmed M. Elgammal, Chan-Su Lee, Separating style and content on a nonlinear manifold, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'04), vol. 1, Washington, DC, June 2004, pp. 478–485.

[29] Markus Enzweiler, Dariu M. Gavrila, Monocular pedestrian detection: survey and experiments, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31(12) (2009) 2179–2195.

[30] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, Xander Twombly, Vision-based hand pose estimation: a review, Computer Vision and Image Understanding (CVIU) 108 (1–2) (2007) 52–73.

[31] María-José Escobar, Guillaume S. Masson, Thierry Vieville, Pierre Kornprobst, Action recognition using a bio-inspired feedforward spiking network, International Journal of Computer Vision (IJCV) 82 (3) (2009) 284–301.

[32] Claudio Fanti, Lihi Zelnik-Manor, Pietro Perona, Hybrid models for human motion recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, San Diego, CA, June 2005, pp. 1166–1173.

[33] Ali Farhadi, Mostafa Kamali Tabriz, Learning to recognize activities from the wrong view point, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 1, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 154–166 (Number 5302).

[34] Ali Farhadi, Mostafa Kamali Tabriz, Ian Endres, David A. Forsyth, A latent model of discriminative aspect, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[35] Alireza Fathi, Greg Mori, Action recognition by learning mid-level motion features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[36] Xiaolin Feng, Pietro Perona, Human action recognition by sequence of movelet codewords, in: Proceedings of the International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'02), Padova, Italy, June 2002, pp. 717–721.

[37] Roman Filipovych, Eraldo Ribeiro, Learning human motion models from unsegmented videos, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–7.

[38] David A. Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien, Deva Ramanan, Computational studies of human motion part 1: tracking and motion synthesis, Foundations and Trends in Computer Graphics and Vision 1 (2) (2006) 77–254.

[39] Yoav Freund, Robert E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Journal of Computer and System Sciences 55 (1) (1997) 119–139.

[40] Adrien Gaidon, Marcin Marszałek, Cordelia Schmid, Mining visual actions from movies, in: Proceedings of the British Machine Vision Conference (BMVC'09), London, United Kingdom, in press.

[41] Tarak Gandhi, Mohan M. Trivedi, Pedestrian protection systems: issues, survey, and challenges, IEEE Transactions On Intelligent Transportation Systems 8 (3) (2007) 413–430.

[42] Dariu M. Gavrila, The visual analysis of human movement: a survey, Computer Vision and Image Understanding (CVIU) 73 (1) (1999) 82–92.

[43] David Gerónimo, Antonio M. López, Angel D. Sappa, Thorsten Graf, Survey of pedestrian detection for advanced driver assistance systems, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), in press.

[44] Andrew Gilbert, John Illingworth, Richard Bowden, Scale invariant action recognition using compound features mined from dense spatio-temporal corners, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 1, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 222–233 (Number 5302).

[45] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, Ronen Basri, Actions as space–time shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (12) (2007) 2247–2253.

[46] Matthias Grundmann, Franziska Meier, Irfan Essa, 3D shape context and distance transform for action recognition, in: Proceedings of the International Conference on Pattern Recognition (ICPR'08), Tampa, FL, December 2008, pp. 1–4.

[47] Abhinav Gupta, Aniruddha Kembhavi, Larry S. Davis, Observing human–object interactions: using spatial and functional compatibility for recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31 (10) (2009) 1775–1789.

[48] Sonal Gupta, Raymond J. Mooney, Using closed captions to train activity recognizers that improve video retrieval, in: Proceedings of the Workshop on Visual and Contextual Learning (VCL) at the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[49] Chris Harris, Mike Stephens, A combined corner and edge detector, in: Proceedings of the Alvey Vision Conference, Manchester, United Kingdom, August 1988, pp. 147–151.

[50] Kardelen Hatun, Pınar Duygulu, Pose sentences: a new representation for action recognition using sequence of pose words, in: Proceedings of the International Conference on Pattern Recognition (ICPR'08), Tampa, FL, December 2008, pp. 1–4.

[51] Yuxiao Hu, Liangliang Cao, Fengjun Lv, Shuicheng Yan, Yihong Gong, Thomas S. Huang, Action detection in complex scenes with spatial and temporal ambiguities, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[52] Feiyue Huang, Guangyou Xu, Viewpoint insensitive action recognition using envelop shape, in: Proceedings of the Asian Conference on Computer Vision (ACCV'07) – part 2, Lecture Notes in Computer Science, Tokyo, Japan, November 2007, pp. 477–486 (Number 4844).

[53] Nazlı İkizler, Ramazan G. Cinbiş, Pınar Duygulu, Human action recognition with line and flow histograms, in: Proceedings of the International Conference on Pattern Recognition (ICPR'08), Tampa, FL, December 2008, pp. 1–4.

[54] Nazlı İkizler, Ramazan G. Cinbiş, Selen Pehlivan, Pınar Duygulu, Recognizing actions from still images, in: Proceedings of the International Conference on Pattern Recognition (ICPR'08), Tampa, FL, December 2008, pp. 1–4.

[55] Nazlı İkizler, Ramazan G. Cinbiş, Stan Sclaroff, Learning actions from the web, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[56] Nazlı İkizler, Pınar Duygulu, Histogram of oriented rectangles: a new pose descriptor for human action recognition, Image and Vision Computing 27 (10) (2009) 1515–1526.

[57] Nazlı İkizler, David A. Forsyth, Searching for complex human activities with no visual examples, International Journal of Computer Vision (IJCV) 30 (3) (2008) 337–357.

[58] Hueihan Jhuang, Thomas Serre, Lior Wolf, Tomaso Poggio, A biologically inspired system for action recognition, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[59] Kui Jia, Dit-Yan Yeung, Human action recognition using local spatio-temporal discriminant embedding, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[60] Hao Jiang, David R. Martin, Finding actions using shape flows, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 2, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 278–292 (Number 5303).

[61] Imran Junejo, Emilie Dexter, Ivan Laptev, Patrick Pérez, Cross-view action recognition from temporal self-similarities, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 2, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 293–306 (Number 5303).

[62] Timor Kadir, Michael Brady, Scale saliency: a novel approach to salient feature and scale selection, in: Proceedings of the International Conference on Visual Information Engineering (VIE), Guildford, United Kingdom, July 2003, pp. 25–28.

[63] Yan Ke, Rahul Sukthankar, Martial Hebert, Efficient visual event detection using volumetric features, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 166–173.

[64] Yan Ke, Rahul Sukthankar, Martial Hebert, Event detection in crowded videos, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[65] Yan Ke, Rahul Sukthankar, Martial Hebert, Spatio-temporal shape and flow correlation for action recognition, in: Proceedings of the Workshop on Visual Surveillance (VS) at the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[66] Vili Kellokumpu, Guoying Zhao, Matti Pietikäinen, Human activity recognition using a dynamic texture based method, in: Proceedings of the British Machine Vision Conference (BMVC'08), Leeds, United Kingdom, September 2008, pp. 885–894.

[67] Tae-Kyun Kim, Roberto Cipolla, Canonical correlation analysis of video volume tensors for action categorization and detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31 (8) (2009) 1415–1428.

[68] Alexander Kläser, Marcin Marszałek, Cordelia Schmid, A spatio-temporal descriptor based on 3d-gradients, in: Proceedings of the British Machine Vision Conference (BMVC'08), Leeds, United Kingdom, September 2008, pp. 995–1004.

[69] Volker Krüger, Danica Kragic, Aleš Ude, Christopher Geib, The meaning of action: a review on action recognition and mapping, Advanced Robotics 21 (13) (2007) 1473–1501.

[70] John D. Lafferty, Andrew McCallum, Fernando C. Pereira, Conditional random fields: probabilistic models for segmenting and labeling sequence data, in: Proceedings of the International Conference on Machine Learning (ICML'01), Williamstown, MA, June 2001, pp. 282–289.

[71] Ivan Laptev, Barbara Caputo, Christian Schüldt, Tony Lindeberg, Local velocity-adapted motion events for spatio-temporal recognition, Computer Vision and Image Understanding (CVIU) 108 (3) (2007) 207–229.

[72] Ivan Laptev, Tony Lindeberg, Space–time interest points, in: Proceedings of the International Conference on Computer Vision (ICCV'03), vol. 1, Nice, France, October 2003, pp. 432–439.

[73] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, Benjamin Rozenfeld, Learning realistic human actions from movies, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[74] Ivan Laptev, Patrick Pérez, Retrieving actions in movies, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[75] Zhe Lin, Zhuolin Jiang, Larry S. Davis, Recognizing actions by shape-motion prototype trees, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[76] Jingen Liu, Saad Ali, Mubarak Shah, Recognizing human actions using multiple features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[77] Jingen Liu, Jiebo LUO, Mubarak Shah, Recognizing realistic actions from videos "in the wild", in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[78] Jingen Liu, Mubarak Shah, Learning human actions via information maximization, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[79] David G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision (IJCV) 60 (2) (2004) 91–110.

[80] Wei-Lwun Lu, James J. Little, Simultaneous tracking and action recognition using the PCA–HOG descriptor, in: Proceedings of the Canadian Conference on Computer and Robot Vision (CRV'06), Quebec City, Canada, June 2006, pp. 6–6.

[81] Fengjun Lv, Ram Nevatia, Recognition and segmentation of 3-D human action using HMM and multi-class adaBoost, in: Proceedings of the European Conference on Computer Vision (ECCV'06), Lecture Notes in Computer Science, vol. 4, Graz, Austria, May 2006, pp. 359–372 (Number 3953).

[82] Fengjun Lv, Ram Nevatia, Single view human action recognition using key pose matching and Viterbi path searching, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[83] Marcin Marszałek, Ivan Laptev, Cordelia Schmid, Actions in context, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[84] Osama Masoud, Nikos Papanikolopoulos, A method for human action recognition, Image and Vision Computing 21 (8) (2003) 729–743.

[85] Pyry Matikainen, Martial Hebert, Rahul Sukthankar, Yan Ke, Fast motion consistency through matrix quantization, in: Proceedings of the British Machine Vision Conference (BMVC'08), Leeds, United Kingdom, September 2008, pp. 1055–1064.

[86] M. Ángeles Mendoza, Nicolás Pérez de la Blanca, Applying space state models in human action recognition: a comparative study, in: International Workshop on Articulated Motion and Deformable Objects (AMDO'08), Lecture Notes in Computer Science, Port d'Andratx, Spain, July 2008, pp. 53–62 (Number 5098).

[87] Ross Messing, Chris Pal, Henry Kautz, Activity recognition using the velocity histories of tracked keypoints, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[88] Krystian Mikolajczyk, Hirofumi Uemura, Action recognition with motion-appearance vocabulary forest, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[89] Sushmita Mitra, Tinku Acharya, Gesture recognition: a survey, IEEE Transactions on Systems, Man, and Cybernetics (SMC) – Part C: Applications and Reviews 37 (3) (2007) 311–324. May.

[90] Thomas B. Moeslund, Adrian Hilton, Volker Krüger, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 90–126.

[91] Darnell J. Moore, Irfan A. Essa, Monson H. Hayes III, Exploiting human actions and object context for recognition tasks, in: Proceedings of the International Conference on Computer Vision (ICCV'99), vol. 1, Kerkyra, Greece, September 1999, pp. 80–86.

[92] Pradeep Natarajan, Ram Nevatia, Online, real-time tracking and recognition of human actions, in: Proceedings of the Workshop on Motion and Video Computing (WMVC'08), Copper Mountain, CO, January 2008, pp. 1–8.

[93] Pradeep Natarajan, Ram Nevatia, View and scale invariant action recognition using multiview shape-flow models, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[94] Juan Carlos Niebles, Li Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[95] Juan Carlos Niebles, Hongcheng Wang, Li Fei-Fei, Unsupervised learning of human action categories using spatial–temporal words, International Journal of Computer Vision (IJCV) 79 (3) (2008) 299–318.

[96] Huazhong Ning, Yuxiao Hu, Thomas S. Huang, Searching human behaviors using spatial–temporal words, in: Proceedings of the International Conference on Image Processing (ICIP'07), vol. 6, San Antonio, TX, September 2007, pp. 337–340.

[97] Huazhong Ning, Wei Xu, Yihong Gong, Thomas S. Huang, Latent pose estimator for continuous action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 2, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 419–433 (Number 5305).

[98] Sebastian Nowozin, Gökhan Bakır, Koji Tsuda, Discriminative subsequence mining for action classification, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[99] Abhijit S. Ogale, Alap Karapurkar, Yiannis Aloimonos, View-invariant modeling and recognition of human actions using grammars, in: Revised Papers of the Workshops on Dynamical Vision (WDV'05 and WDV'06), Lecture Notes in Computer Science, Beijing, China, May 2007, pp. 115–126 (Number 4358).

[100] Takehito Ogata, William Christmas, Josef Kittler, Seiji Ishikawa, Improving human activity detection by combining multi-dimensional motion descriptors with boosting, in: Proceedings of the International Conference on Pattern Recognition (ICPR'06), vol. 1, Kowloon Tong, Hong Kong, August 2006, pp. 295–298.

[101] Antonios Oikonomopoulos, Maja Pantic, Ioannis Patras, Sparse B-spline polynomial descriptors for human activity recognition, Image and Vision Computing 27(12) (2009) 1814–1825.

[102] Antonios Oikonomopoulos, Ioannis Patras, Maja Pantic, Spatiotemporal salient points for visual recognition of human actions, IEEE Transactions On Systems Man And Cybernetics (SMC) – Part B: Cybernetics 36 (3) (2006) 710–719.

[103] Olusegun Oshin, Andrew Gilbert, John Illingworth, Richard Bowden, Spatio-temporal feature recognition using randomised ferns, in: Proceedings of the International Workshop on Machine Learning for Vision-based Motion Analysis (MLVMA'08), Marseille, France, October 2008, pp. 1–12.

[104] Vasu Parameswaran, Rama Chellappa, View invariance for human action recognition, International Journal of Computer Vision (IJCV) 66 (1) (2006) 83–101.

[105] Sangho Park, Mohan M. Trivedi, Understanding human interactions with track and body synergies (TBS) captured from multiple views, Computer Vision and Image Understanding (CVIU) 111 (1) (2008) 2–20.

[106] Alonso Patron-Perez, Ian Reid, A probabilistic framework for recognizing similar actions using spatio-temporal features, in: Proceedings of the British Machine Vision Conference (BMVC'07), Edinburgh, United Kingdom, September 2007, pp. 1–10.

[107] Patrick Peursum, Svetha Venkatesh, Geoff West, Tracking-as-recognition for articulated full-body human motion analysis, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[108] Ramprasad Polana, Randal C. Nelson, Detection and recognition of periodic, nonrigid motion, International Journal of Computer Vision (IJCV) 23 (3) (1997) 261–282.

[109] Ronald Poppe, Vision-based human motion analysis: an overview, Computer Vision and Image Understanding (CVIU) 108 (1-2) (2007) 4–18.

[110] Ronald Poppe, Mannes Poel, Discriminative human action recognition using pairwise CSP classifiers, in: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FGR'08), September 2008, Amsterdam, The Netherlands, 2008, pp. 1–6.

[111] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, Trevor Darrell, Hidden conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (10) (2007) 1848–1852.

[112] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, Tim Ellis, Human action recognition using robust power spectrum features, in: Proceedings of the International Conference on Image Processing (ICIP'08), San Diego, CA, October 2008, pp. 753–756.

[113] Deva Ramanan, Learning to parse images of articulated bodies, in: Advances in Neural Information Processing Systems (NIPS), vol. 19, Vancouver, Canada, December 2006, pp. 1129–1136.

[114] Deva Ramanan, David A. Forsyth, Automatic annotation of everyday movements, in: Advances in Neural Information Processing Systems (NIPS), vol. 16, Vancouver, Canada, 2003, pp. 1–8.

[115] Deva Ramanan, David A. Forsyth, Andrew Zisserman, Tracking people by learning their appearance, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (1) (2007) 65–81.

[116] Cen Rao, Alper Yilmaz, Mubarak Shah, View-invariant representation and recognition of actions, International Journal of Computer Vision (IJCV) 50 (2) (2002) 203–226.

[117] Konstantinos Rapantzikos, Yannis Avrithis, Stefanos Kollias, Dense saliency-based spatiotemporal feature points for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[118] Konstantinos Rapantzikos, Yannis S. Avrithis, Stefanos D. Kollias, Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: potential in human action recognition, in: Proceedings of the International Conference on Image and Video Retrieval (CIVR'07), July 2007, Amsterdam, The Netherlands, 2007, pp. 294–301.

[119] Neil Robertson, Ian Reid, A general method for human activity recognition in video, Computer Vision and Image Understanding (CVIU) 104 (2) (2006) 232–248.

[120] Mikel D. Rodriguez, Javed Ahmed, Mubarak Shah, Action MACH: a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[121] Rómer E. Rosales, Recognition of human action using moment-based features, Technical Report BU-1998-020, Boston University, Computer Science, Boston, MA, November 1998.

[122] Michael S. Ryoo, Jake K. Aggarwal, Semantic representation and recognition of continued and recursive human activities, International Journal of Computer Vision (IJCV) 82 (1) (2009) 1–24.

[123] Silvio Savarese, Andrey DelPozo, Juan Carlos Niebles, Li Fei-Fei, Spatial–temporal correlatons for unsupervised action classification, in: Proceedings of the Workshop on Applications of Computer Vision (WACV'08), Copper Mountain, CO, January 2008, pp. 1–8.

[124] Konrad Schindler, Luc J. van Gool, Action snippets: how many frames does human action recognition require? in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[125] Christian Schüldt, Ivan Laptev, Barbara Caputo, Recognizing human actions: a local SVM approach, Proceedings of the International Conference on Pattern Recognition (ICPR'04), 2004, vol. 3, Cambridge, United Kingdom, 2004, pp. 32–36.

[126] Paul Scovanner, Saad Ali, Mubarak Shah, A 3-dimensional SIFT descriptor and its application to action recognition, in: Proceedings of the International Conference on Multimedia (MultiMedia'07), Augsburg, Germany, September 2007, pp. 357–360.

[127] Steven M. Seitz, Charles R. Dyer, View-invariant analysis of cyclic motion, International Journal of Computer Vision (IJCV) 25 (3) (1997) 231–251.

[128] Hae Jong Seo, Peyman Milanfar, Detection of human actions from a single example, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[129] Eli Shechtman, Michal Irani, Matching local self-similarities across images and videos, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[130] Eli Shechtman, Michal Irani, Space–time behavior-based correlation-OR-How to tell if two underlying motion fields are similar without computing them?, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 29 (11) (2007) 2045–2056

[131] Yaser Sheikh, Mumtaz Sheikh, Mubarak Shah, Exploring the space of a human action, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 144–149.

[132] Yuping Shen, Hassan Foroosh, View-invariant action recognition from point triplets, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31 (10) (2009) 1898–1905.

[133] Qinfeng Shi, Li Wang, Li Cheng, Alex Smola, Discriminative human action segmentation and recognition using semi-Markov model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[134] Cristian Sminchisescu, Atul Kanaujia, Dimitris N. Metaxas, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 210–220.

[135] Paul Smith, Niels da Vitoria Lobo, Mubarak Shah, TemporalBoost for event recognition, in: Proceedings of the International Conference On Computer Vision (ICCV'05), vol. 1, Beijing, China, October 2005, pp. 733–740.

[136] Yang Song, Luis Goncalves, Pietro Perona, Unsupervised learning of human motion, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 25 (7) (2003) 814–827.

[137] Richard Souvenir, Justin Babbs, Learning the viewpoint manifold for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–7.

[138] Josephine Sullivan, Stefan Carlsson, Recognizing and tracking human action, in: Proceedings of the European Conference on Computer Vision (ECCV'02), vol. 1, Lecture Notes in Computer Science, Copenhagen, Denmark, May 2002, pp. 629–644 (Number 2350).

[139] Evan A. Suma, Christopher W. Sinclair, Justin Babbs, Richard Souvenir, A sketch-based approach for detecting common human actions, in: Proceedings of the International Symposium on Advances in Visual Computing (ISVC'08) – part 1, Lecture Notes in Computer Science, Las Vegas, NV, December 2008, pp. 418–427 (Number 5358).

[140] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua, Jintao Li, Hierarchical spatio-temporal context modeling for action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[141] Christian Thurau, Václav Hlaváč, Pose primitive based human action recognition in videos or still images, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–6.

[142] Du Tran, Alexander Sorokin, David A. Forsyth, Human activity recognition with metric learning, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 1, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 548–561 (Number 5302).

[143] Pavan Turaga, Rama Chellappa, Venkatramana S. Subrahmanian, Octavian Udrea, Machine recognition of human activities: a survey, IEEE Transactions on Circuits and Systems for Video Technology 18 (11) (2008) 1473–1488.

[144] Pavan Turaga, Ashok Veeraraghavan, Rama Chellappa, Unsupervised view and rate invariant clustering of video sequences, Computer Vision and Image Understanding (CVIU) 113 (3) (2009) 353–371.

[145] Pavan K. Turaga, Ashok Veeraraghavan, Rama Chellappa, Statistical analysis on stiefel and grassmann manifolds with applications in computer vision, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[146] Hirofumi Uemura, Seiji Ishikawa, Krystian Mikolajczyk, Feature tracking and motion compensation for action recognition, in: Proceedings of the British Machine Vision Conference (BMVC'08), Leeds, United Kingdom, September 2008, pp. 293–302.

[147] Ashok Veeraraghavan, Rama Chellappa, Amit K. Roy-Chowdhury, The function space of an activity, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 1, New York, NY, June 2006, pp. 959–968.

[148] Ashok Veeraraghavan, Amit K. Roy-Chowdhury, Rama Chellappa, Matching shape sequences in video with applications in human movement analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 27 (12) (2005) 1896–1909.

[149] Paul A. Viola, Michael J. Jones, Rapid object detection using a boosted cascade of simple features, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01), vol. 1, Kauai, HI, December 2001, pp. 511–518.

[150] Shiv N. Vitaladevuni, Vili Kellokumpu, Larry S. Davis, Action recognition using ballistic dynamics, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–8.

[151] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, Cordelia Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of the British Machine Vision Conference (BMVC'09), London, United Kingdom, in press.

[152] Jack M. Wang, David J. Fleet, Aaron Hertzmann, Multifactor Gaussian process models for style-content separation, in: Proceedings of the International Conference on Machine Learning (ICML'07), ACM International Conference Proceeding Series, Corvalis, OR, June 2007, pp. 975–982 (Number 227).

[153] Liang Wang, Weiming Hu, Tieniu Tan, Recent developments in human motion analysis, Pattern Recognition 36 (3) (2003) 585–601.

[154] Liang Wang, David Suter, Informative shape representations for human action recognition, in: Proceedings of the International Conference on Pattern Recognition (ICPR'06), vol. 2, Kowloon Tong, Hong Kong, August 2006, pp. 1266–1269.

[155] Liang Wang, David Suter, Learning and matching of dynamic shape manifolds for human action recognition, IEEE Transactions On Image Processing (TIP) 16 (6) (2007) 1646–1661.

[156] Liang Wang, David Suter, Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[157] Liang Wang, David Suter, Visual learning and recognition of sequential data manifolds with applications to human movement analysis, Computer Vision and Image Understanding (CVIU) 110 (2) (2008) 153–172.

[158] Yang Wang, Hao Jiang, Mark S. Drew, Ze-Nian Li, Greg Mori, Unsupervised discovery of action classes, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, New York, NY, June 2006, pp. 1654–1661.

[159] Yang Wang, Greg Mori, Learning a discriminative hidden part model for human action recognition, in: Advances in Neural Information Processing Systems (NIPS), vol. 21, Vancouver, Canada, December 2008, pp. 1721–1728.

[160] Yang Wang, Greg Mori, Human action recognition by semilatent topic models, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 31 (10) (2009) 1762–1774. October.

[161] Yang Wang, Greg Mori, Max-margin hidden conditional random fields for human action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[162] Ying Wang, Kaiqi Huang, Tieniu Tan, Human activity recognition based on $\mathfrak{R}$ transform, in: Proceedings of the Workshop on Visual Surveillance (VS) at the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[163] Daniel Weinland, Edmond Boyer, Action recognition using exemplar-based embedding, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–7.

[164] Daniel Weinland, Edmond Boyer, Remi Ronfard, Action recognition from arbitrary views using 3D exemplars, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[165] Daniel Weinland, Remi Ronfard, Edmond Boyer, Automatic discovery of action taxonomies from multiple views, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, New York, NY, June 2006, pp. 1639–1645.

[166] Daniel Weinland, Remi Ronfard, Edmond Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 249–257.

[167] Geert Willems, Tinne Tuytelaars, Luc J. Van Gool, An efficient dense and scale-invariant spatio-temporal interest point detector, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 2, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 650–663 (Number 5303).

[168] Shu-Fai Wong, Roberto Cipolla, Extracting spatiotemporal interest points using global information, in: Proceedings of the International Conference On Computer Vision (ICCV'07), Rio de Janeiro, Brazil, October 2007, pp. 1–8.

[169] Shu-Fai Wong, Tae-Kyun Kim, Roberto Cipolla, Learning motion categories using both semantic and structural information, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, June 2007, pp. 1–8.

[170] Junji Yamato, Jun Ohya, Kenichiro Ishii, Recognizing human action in time-sequential images using hidden Markov model, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'92), Champaign, IL, June 1992, pp. 379–385.

[171] Pingkun Yan, Saad M. Khan, Mubarak Shah, Learning 4D action feature models for arbitrary view action recognition, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'08), Anchorage, AK, June 2008, pp. 1–7.

[172] Changjiang Yang, Yanlin Guo, Harpreet S. Sawhney, Rakesh Kumar, Learning actions using robust string kernels, in: Human Motion: Understanding, Modeling, Capture and Animation (HUMO'07), Lecture Notes in Computer Science, Rio de Janeiro, Brazil, October 2007, pp. 313–327 (Number 4814).

[173] Benjamin Yao, Song-Chun Zhu, Learning deformable action templates from cluttered videos, in: Proceedings of the International Conference On Computer Vision (ICCV'09), Kyoto, Japan, September 2009, pp. 1–8.

[174] Alper Yilmaz, Mubarak Shah, Matching actions in presence of camera motion, Computer Vision and Image Understanding (CVIU) 104 (2–3) (2006) 221–231.

[175] Alper Yilmaz, Mubarak Shah, A differential geometric approach to representing the human actions, Computer Vision and Image Understanding (CVIU) 119 (3) (2008) 335–351.

[176] Junsong Yuan, Zicheng Liu, Ying Wu, Discriminative subvolume search for efficient action detection, in: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'09), Miami, FL, June 2009, pp. 1–8.

[177] Lihi Zelnik-Manor, Michal Irani, Statistical analysis of dynamic actions, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 28 (9) (2006) 1530–1535.

[178] Jianguo Zhang, Shaogang Gong, Action categorization with modified hidden conditional random field, Pattern Recognition 43 (1) (2010) 197–203.

[179] Ziming Zhang, Yiqun Hu, Syin Chan, Liang-Tien Chia, Motion context: a new representation for human action recognition, in: Proceedings of the European Conference on Computer Vision (ECCV'08) – part 4, Lecture Notes in Computer Science, Marseille, France, October 2008, pp. 817–829 (Number 5305).

[180] Zhipeng Zhao, Ahmed Elgammal, Human activity recognition from frame's spatiotemporal representation, in: Proceedings of the International Conference on Pattern Recognition (ICPR'08), Tampa, FL, December 2008, pp. 1–4.