

Chapter 2

Action Representation

Abstract In this chapter, various action recognition issues are covered in a concise manner. Various approaches are presented here. In Chap. 1, nomenclatures, various aspects of action recognition are detailed. Hence, the previous chapter is crucial to provide a base for this chapter.

2.1 Action Recognition

Action recognition is a very important area in computer vision and other fields. We have different approaches for action recognition. These are defined based on the characteristics of different methods and their inherent strategies to deal with action classes. In one of the dominant classifications, action recognition approaches are divided into the following categories [505, 567]:

- Template matching approaches.
- State-space approaches.
- Semantic description of human behaviors.

Before presenting the Motion History Image method, we need to know the diversified approaches for action recognitions. Therefore, we devote this chapter to classify major approaches in various dimensions. We will notice that a number of approaches have been explored to solve the core problems of action recognition. Our concentration is mainly on the action *representations*—because without having smart and robust representations, it will not be possible to recognize actions in a reasonable manner. Though the pattern recognition part is significantly crucial for action recognition, this book does not deal much with machine learning methodologies. We feel that there are already a good amount of literatures in that arena.

Although it is a fact that researchers are investing enormous efforts to propose and justify their approaches for action or activity recognition—the reality is that this field is still not mature enough to be applicable in many important areas. As [558] reported earlier—the activity representation and recognition is relatively old,

yet still immature! Different dimensions can be achieved from some survey papers on action recognition and related issues [494, 505, 506, 510, 539, 546–565, 700]. The following sections present different approaches for action recognition.

2.2 Approaches on Bag-of-Features

The bag-of-features approach [64] is a well-known method for action recognition. The *bag-of-features*-based approaches can be applied in classification by employing features as *words*. Due to its popularity, researchers are extensively considering this framework for their researches [148, 168, 440, 441, 446, 448, 449, 498, 601, 637, 638, 643, 646, 651, 652, 657]. It has similar versions in the literature as:

- Bag-of-Features [64, 168, 638, 652]
- Bag-of-Words [168]
- Bag-of-Visual-Words [168]
- Bag-of-Vocabularies [439]
- Bag-of-Video-Words [168]
- Bag-of-Points [697]

Now, what is the meaning of *bag* here as well as *words/features*? In fact, in order to represent an image template by considering the framework of bag-of-features or bag-of-words or similar—one can consider that image as a *document*, where *words* can be produced from the following steps:

- **Step-1—Feature detection:**

This is one of the elementary image processing steps where image patches or points are analyzed to retrieve any significant image features. Now, what constitutes *features*? There is in fact no specific boundary or threshold to define a *feature* from an image template. A feature can be treated as the *interested points* for that specific image, which might not be *interesting* for another image or application. *Good features* are not defined and these are application-specific.

- **Step-2—Feature description or feature representation:**

Once we have the image features—detected by one of the above-mentioned approaches, we can compute *numerical vectors* from these features. So, from an image—we extract features—then compute feature vectors by using an approach for feature description, called *feature descriptors*.

- **Step-3—Discrete vocabularies or dictionary or codebook generation:**

From the above stage, we can have feature vectors from each image template. The feature vectors have equal dimensions for each image though the orders of different vectors have no importance. It is a constraint of bag-of-features. In the final stage, features are quantized into *discrete vocabularies* or *codewords* (similar to *words* in a *document*) and hence a *codebook* (similar to a *word dictionary*) are produced. These are clustered and hence a suitable clustering method should be selected. Therefore, we have a certain codeword that can relate to an image patch, and the image is represented by a histogram of the codewords.

The Bag-of-Features representation is typically a normalized histogram, where each bin in the histogram is the number of features assigned to a particular code divided by the total number of features in the video clip [709]. In short, a bag-of-words approach is as follows:

- Generates a vocabulary of visual words.
- Characterizes videos with the histograms of visual word counts.

In case of a video, we need to find a suitable approach to sample a video to extract localized features. Some approaches are [439]:

- Space-time interest point operators
- Grids/pyramids
- Random sampling

There are some clear-cut advantages of the bag-of-words framework, as [439]:

- It presents simple representation.
- It needs almost no preprocessing steps.

Some disadvantages of this framework are:

- We can notice in the steps of bag-of-features paradigm that the entire spatial arrangements of features are lost. This may be a major problem in some applications where relationships among various spatial arrangements are necessary. It is not possible to have an explicit model of a subject and hence, it cannot provide localized information.
- Another constraint is related to the missing spatio-temporal information. Thus, in case of any actions with several motion or posture changes (e.g., opening a door) and symmetric actions (e.g., ‘sitting down’ versus ‘standing up’ actions), this framework cannot perform well [150].
- In the recognition process, this framework cannot explicitly justify on where two actions are matching each other [439].
- If the codebook becomes very large, it may produce lower recognition. On the other hand, if the vocabulary size is small, it may cause over-clustering and poor recognition.
- The feature detection and vocabulary generation is time-taxing for large amount of data. Moreover, prior heavy training loads may deter the performance [632].
- Another constraint is related with the necessity of knowing the number of visual features, the size of vocabularies, the level of hierarchies, and the number of kernels for clustering, etc.

Some approaches to improve the performance of bag-of-words approaches:

Due to its simplicity—the bag-of-features or bag-of-words frameworks get wide attention among researchers—not only for action recognition but also for object classification. Recently, a few approaches have been proposed to mitigate some of these constraints. For example, a bag-of-features method is proposed for learning

discriminative features on space-time neighborhoods by [637], where the descriptors are modeled hierarchically and multiple kernel learning is used to characterize each action.

Usually, large vocabulary size of the bag-of-visual-words is more discriminative for inter-class action classification while a small one is more robust to noise and thus tolerant to the intra-class invariance [446]. In other words, it is common to choose an appropriately large vocabulary size. The larger the vocabulary size, the more chance to have a sparse histogram for each video, and thereby yield more noise and reduce the discriminability of vocabulary.

On the other side, if the vocabulary size is small, it may cause over-clustering and high intra-class distortion. In order to overcome this shortcoming, a pyramid vocabulary tree is proposed [446]. They cluster the interest points in the spatio-temporal space. The spatio-temporal space forms some cluster centers, where histograms of local features are produced. A sparse spatio-temporal pyramid matching kernel (SST-PMK) is proposed to compute the similarity measures between video sequences [446]. SST-PMK satisfies the Mercers condition and therefore is readily integrated into SVM to perform action recognition. They found that both the pyramid vocabulary tree and the SST-PMK lead to a significant improvement in human action recognition on the Weizmann dataset.

2.3 XYT: Space-Time Volume

Action recognition approaches can be based on space-time volume or spatio-temporal features. The *spatio* term is related to the *XY*/spatial domain and the *temporal* term is noting the *T*/time of an action. Let us consider an action in a video clip. We want to have its representation as a spatio-temporal template by combining the spatial motion information along with the temporal information. And from this template—if we can achieve significant motion information along its motion duration, then it will be very significant information for action recognition. This process is attempted to propose space-time volume-based methods for action recognition. One of the simplest but effective methods is the Motion History Image (MHI) method [492]. The MHI method itself consumes the temporal information in a template or final image from a video scene. Apart from the MHI method, a number of other methods have combined or incorporated the spatio-temporal information for better action representations.

Some key characteristics of the spatio-temporal features are:

- Space(*X,Y*)-time(*T*) descriptors may strongly depend on the relative motion between the object and camera.
- Some corner points in time, called space-time interest points, can automatically adapt the features to the local velocity of the image pattern.
- However, these space-time points are often found on highlights and shadows, hence, they are sensitive to lighting conditions and it may affect recognition accuracy.



Fig. 2.1 Example of spatio-temporal interest points (STIP) for an action

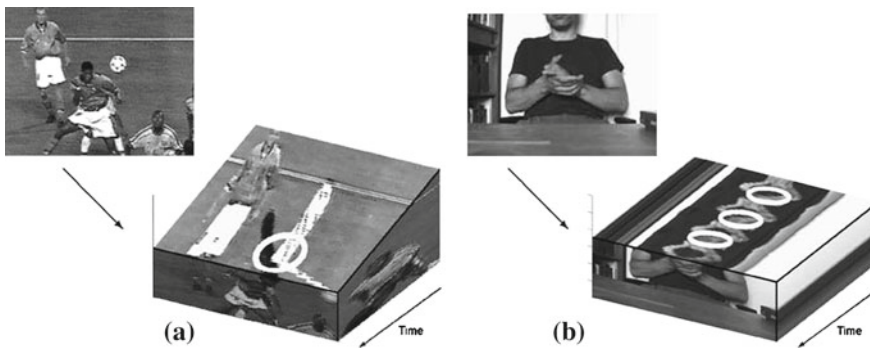


Fig. 2.2 Examples of detecting the strongest spatio-temporal interest points (STIP) **a** a football sequence with a player heading a ball; **b** a hand clapping sequence. With kind permission of Springer Science + Business Media B.V.—from Laptev [440]: Fig. 1, Springer, 2005

- Spatio-temporal features can avoid some limitations of traditional approaches of intensities, gradients, optical flow, and other local features.

Figure 2.1 depicts some spatio-temporal interest points for a walking action. Figure 2.2 shows some results of detecting the strongest spatio-temporal interest points in a football sequence with a player heading the ball and in a hand clapping sequence [440]. From these temporal slices of space-time volumes, it is evident that the detected events correspond to neighborhoods with high spatio-temporal variation in the image data.

Now we present a few methods that employ the spatio-temporal concepts to represent and extract motion information for recognition. References [600, 601] propose a volumetric *space-time action shapes* that are induced by a concatenation of 2D silhouettes in the space-time volume and contain both the spatial information about the posture of the human subject at any time (location and orientation of the torso and

limbs, aspect ratio of different body parts), as well as the dynamic information (global body motion and motion of the limbs relative to the body) [601]. In the approach by [600], each internal point is assigned with the mean time required for a particle undergoing a random-walk process starting from the point to hit the boundaries. This method utilizes properties of the solution to the Poisson equation to extract space-time features such as local space-time saliency, action dynamics, shape structure, and orientation. These features are useful for action recognition, detection, and clustering [601]. In another similar approach, human actions are presented as 3D spatio-temporal surfaces and analyzed using differential geometric surface properties [608].

In Chap. 3, we present the MHI method and its variants in detail and most of the approaches are proposed with the concept of space-time volume, as these representations incorporate cumulative temporal information within the spatial domain.

2.3.1 Spatio-Temporal Silhouettes

Spatio-temporal silhouettes are exploited by various researches [483, 490, 531, 602, 608, 612, 614, 645]. From a video stream, binary silhouettes are extracted by some means and then these are used to produce action representations. Some of the approaches based on spatio-temporal silhouettes are presented below.

The *Silhouette History Image* (SHI) and the *Silhouette Energy Image* (SEI) [309] are proposed that exploit silhouettes for these two representations. The successive silhouette differences are set in the MHI in such a manner that motion from the silhouette boundary can be perceived in the gradient of the MHI [285]. The *timed Motion History Image* (tMHI) is proposed by Bradski and Davis [369] to generalize the MHI to directly encode time in a floating-point format. Another silhouette-based action modeling for recognition is presented by [324]. A probabilistic graphical model using ensembles of spatio-temporal patches is proposed by [614] in order to detect irregular behaviors in videos.

A number of gait recognition methods are based on spatio-temporal silhouettes. For example, the *Gait Energy Image* (GEI) [44, 297, 298], the *Frame Difference Energy Image* (FDEI) [42], the *Frame Difference History Image* (FDHI) [41], the *Average Motion Energy* (AME), the *Mean Motion Shape* (MMS) [612], the *Volume Motion Template* (VMT) [276, 502], the *Volumetric Motion History Image* (VMHI) [291, 328], the *Motion History Volume* (MHV) [531], and other methods [483, 490].

2.4 Interest-Point Detectors

Interest point detection is key to many methods for action recognition. For an image, finding the appropriate interest points and detecting those points or features are one part—while, for consecutive images from a video sequence, describing these points

by feature descriptors is another dimension. One of the earliest but well-known and well-employed spatio-temporal feature or interest point detectors is the *Harris-Laplace detector* [440]. Which points or features are the key is difficult to decide [474]. Feature points can be considered based on corner points [440], the presence of global texture [473], periodicity [474], motion flow vectors [498], etc.

As defined above, from an image, features are extracted through some feature descriptors. Some of the important feature descriptors are:

- Scale-Invariant Feature Transform (SIFT).
- Rank Scale-Invariant Feature Transform.
- PCA-SIFT—It is a variant of SIFT.
- Generalized Robust Invariant Feature (G-RIF)—It considers edge density, edge orientation, etc., to produce a generalized robust feature descriptor as an extension of SIFT.
- Rotation-Invariant Feature Transform (RIFT)—It is a rotation-invariant version of SIFT, developed based on circular normalized patches.
- Speeded-Up Robust Feature (SURF), and its variants.
- Gradient Location and Orientation Histogram (GLOH)—It is a SIFT-like descriptor and it considers more spatial regions for the histograms. It is proposed to enhance robustness and distinctiveness of the SIFT.
- Local Energy-based Shape Histogram (LESH)—It is based on local energy model of features.
- Histogram of Oriented Gradients (HOG), and its variants.

The SURF (Speeded-Up Robust Features [399, 475]) is developed for interest point detection. The best feature descriptor should be invariant to rotation (RIFT), scale (e.g., SIFT, LESH), intensity, and affine variations to a significant level. However, none of the above descriptors are invariant to rotation, scale, intensity, and affine changes in a smart manner. Among these, the Scale-Invariant Feature Transform (SIFT) performs well. Now, regarding the computation of *feature descriptor* as numerical vectors, the SIFT can convert an image patch into 128-dimensional vector, SURF can produce 64-dimensional vector (in order to reduce the time for feature computation); PCA-SIFT reduces to 36-dimensions with PCA, GLOH estimates a vector having 128- or 64-dimensions.

There are numerous ways to detect features [477, 618] (e.g., edges, corners, patches, interest points, blobs or region of interest (ROI), ridges, etc.) from an image, such as:

- *Edge detection*
 - Canny edge detection
 - Sobel operator
 - Prewitt filter
 - Roberts Cross
 - Canny-Deriche
 - Differential edge detection

- Hough transforms
- Harris and Stephens / Plessey
- Smallest Univalued Segment Assimilating Nucleus (SUSAN)
- *Blob detection*
 - Laplacian of Gaussian (LoG)
 - Difference of Gaussian (DoG)
 - Determinant of Hessian (DoH)
 - Principal Curvature-based Region Detector (PCBR)
 - Hessian-affine
 - Hessian-Laplace
 - Harris-affine
 - Maximally Stable extremal Regions (MSER)
 - Lindeberg's Watershed-based Gray-level Blob detector
- *Corner detection*
 - Multi-scale Harris operator
 - Harris and Stephens algorithm
 - Level curve curvature approach
 - Affine-adapted interest point operators
 - SUSAN corner detector
 - Wang and Bradley corner detector
 - Accelerated Segment Test (AST)-based feature detector
 - Features from Accelerated Segment Test (FAST)
 - Shi and Tomasi
 - LoG
 - DoG
 - DoH
 - Moravec algorithm
 - Foerstner corner detector
- *Hough transform, Kernel-based Hough transform*
- *Structure tensor*
- *Ridge detector*

These are based on whether regions of the image or the blob or corner should be detected as key interest points. Usually, it is beneficial to exploit spatio-temporal features due to the fact that it may overcome the constraints related to the computation of optical flow, feature tracking, selection of key frames, extraction of silhouettes, etc. [486, 598, 599, 601]. The optical flow is noisy and it faces aperture problems, smooth surfaces, singularities, etc. Feature tracking has the problem of occlusions, self-occlusions, re-initialization, and change of appearance. Interest points can be considered as the salient points or regions. Approaches that deal with these issues are [601, 602, 605, 607, 608, 638, 643, 649, 651, 653, 654].

The following descriptors can encode the spatio-temporal support region of these interest points, according to [36]:

- Histograms of Oriented Gradient (HOG) descriptor
- SIFT and SURF descriptors
- Histograms of Optic Flow (HOF) descriptor
- Point-trajectory features
- Vector of concatenated pixel gradients
- Local jet descriptors
- Volumetric features

Reference [282] propose the *Histograms of Oriented Gradient* (HOG) descriptor [63, 66, 148, 282, 638, 698] for human detection and recognition. The *Histograms of Optic Flow* (HOF) descriptors [63, 148, 638] are proposed based on optical flow. Performances and popularities of the HOF descriptor are less than the HOG descriptors. The generalized SURF and SIFT descriptors are also popular [592, 635, 656, 658]. On the contrary, the descriptors based on point-trajectory features [65, 148], pixel gradients [652], local jet [651] and volumetric features [148, 531] are emerging descriptors and hence, detailed information about their performances and constraints are not sufficient. However, the volumetric features can be useful in view-invariant action recognition.

2.5 Local Discriminative Approaches

In different domains, action recognition is accomplished by:

- Action recognition based on large-scale features.
- Action recognition based on local patches.
- Action recognition based on mixed or mid-level approach of the above two methodologies.

2.5.1 Large-Scale Features-Based Recognition

One of the large-scale features is optical flow-based approach. It is used as a spatio-temporal descriptor for action recognition [182, 493]. Large-scale feature describes the entire human figure. Reference [493] recognizes some actions at a distance where the resolution of each individual is small (e.g., about 30 pixels tall only). A motion descriptor is introduced where, the raw optical flow vector is split into four different channels. Initially, the optical flow vector field is split into two scalar fields corresponding to the horizontal and vertical components of the flow, F_x and F_y , each of which is then half-wave rectified into four non-negative channels [493]. These are called spatio-temporal motion descriptors, which can be considered as large-scale features, encompassing the entire image. They classify three datasets: ballet dataset, real tennis data from a static camera, and soccer game from moving camera.

2.5.2 Local Patches-Based Recognition

Reference [613] classifies spatial histogram feature vectors into prototypes. Local patches mean part-based models for action representation. Local features describe small patches. Though conceptually appealing and promising—the merit of part-based models has not yet been widely recognized in action recognition [182]. A bag-of-words representation can be used to model these local patches for action recognition. One of the key concerns of this approach is that it suffers from the same restriction of conditional independence assumption that ignores the spatial structure of the parts.

2.5.3 Mixed Approach for Recognition

This hybrid approach combines the large-scale features with local patches for better recognition [27, 182, 452]. Reference [182] introduces a similar concept where a human action is represented by the combination of large-scale global features and local patch features. Reference [452] presents a method of recognition that learns mid-level motion features. In another approach, optical flow is used as large-scale features and SURF is used, where extracted SURF descriptors represent local appearances around interest points [27].

If the complexity of actions increases, the recognition methods become more difficult. Therefore, use of a combination of feature types is necessary. For example, with the *Coffee and Cigarettes Dataset*, which is a difficult dataset, a combination of feature types is considered by [659].

2.6 View-Invariant Approaches

View-invariant methods are difficult and hence, most of the methods are view-variant instead of being view-invariant. The Motion History Image method is basically a view-based method. We will see in a later chapter that some approaches have been proposed to convert the MHI from view-based to view-invariant method, namely the Motion History Volumes (MHV), Volume Motion Templates (VMT), 3D-MHI, etc.

It is very difficult to develop a system that is fully view-invariant due to various reasons [36], such as:

- Occlusion due to object, body parts, motion self-occlusion.
- Similar actions seen from different angles may appear as different actions.
- Multi-camera coordination and management.
- Computational cost.

A number of view-invariant methods are available in [21, 22, 24–26, 276, 277, 279, 280, 291, 308, 328, 420, 421, 502, 503, 509, 511, 531, 583, 590, 593, 594]. Most of the

view-invariant methods exploit video streams from multiple-camera from different angles. INRIA IXMAS action dataset is one of the widely used datasets for this purpose. In a few cases, stereo cameras, Kinect sensor (which has stereo camera too) are used [26, 502].

Each action is represented as a unique curve in a 3D invariance-space, surrounded by an acceptance volume called *action-volume*. It is proposed by [23]. Reference [24] proposes a *Spatio-Temporal Manifold* (STM) model to analyze nonlinear multivariate time series with latent spatial structure and applies it to recognize actions in the joint-trajectories space. It is a view-invariant approach. Reference [25] presents a two-layer classification model for view-invariant human action recognition based on interest points. In this work, training videos of every action are recorded from multiple viewpoints and represented as space-time interest points.

Reference [22] presents a framework for learning a compact representation of primitive actions (e.g., walk, punch, kick, sit) that can be used for video obtained from a single camera for simultaneous action recognition and viewpoint estimation. Reference [21] introduces the idea that the motion of an articulated body can be decomposed into rigid motions of planes defined by triplets of body points.

Some view-invariant methods of the 2D Motion History Image (MHI) representation are proposed and these are detailed in Chap. 3. For example, the *Volume Motion Template* (VMT) is proposed as view-invariant 3D action recognition method [502] and they use only a stereo-camera to produce their results in a view-invariance manner. Weinland et al. [279, 503, 531] introduce the *Motion History Volume* (MHV) method—a view-invariant approach with the concept of the 2D MHI. Reference [303] proposes the *3D Motion History Model* (3D-MHM) as a view-invariant MHI. In a similar fashion, [280, 509] incorporate the information of position of body limbs, and develop a method called *Motion History Volume* (MHV) as a 3D-MHI and *Motion Energy Volume* (MEV) as a 3D-MEI template. Another 3D-MHI representation, called the *Volumetric Motion History Image* (VMHI) is proposed by [291, 328].

In another dimension, [26] presents a novel approach for human action recognition with *Histograms of 3D Joint* locations (HOJ3D) as a compact representation of postures. They extract the 3D skeletal joint locations from depth maps from Kinect sensor by using Shotton’s method. As the Kinect sensor is an important addition that provides depth images along with color images—in future, we will see more approaches that lead us to better view-invariant methods.

2.7 Conclusion

In this chapter, we present important divisions and varieties of action recognition strategies. As the book is about the Motion History Image (MHI)—we do not have enough scope to detail other methods. These glimpses provide a foundation on dimensions of action recognition. In Chap. 3, we present the MHI method and its representations in detail.

Motion History Images for Action Recognition and
Understanding

Ahad, M.A.R.

2013, XVI, 116 p. 34 illus., Softcover

ISBN: 978-1-4471-4729-9