HUMAN ACTION RECOGNITION IN VIDEOS: A SURVEY

Farhood Negin, François Bremond

INRIA - Sophia Antipolis
2004 Route de Lucioles, Sophia Antipolis, France

Abstract

This document provides a brief overview of recent developments in human action recognition in videos. In recent years, there is a increasing interest over analyses of human action and activities on videos because of its broad range of applicability. While video retrieval from various datasets (like YouTube) is one of the favorite studies of action recognition, video surveillance, rehabilitation tasks, abnormal activity detection among many others can benefit from it. In general action recognition approaches can be classified in five main categories: Model-based or feature-based. Also, there are lots of activity recognition methods in the literature. Here, we investigate some of major works which have been done in the field over last twenty years, with an emphasis on the recent state-of-the-art developments on each of the topics.

1. Introduction

Over the last years various techniques have been proposed for action recognition in videos. In this document, we divide the existing techniques into these main categories:

- **Hand-Crafted, knowledge based activity recognition** (Section 2) rather than focusing on a particular action, it tries to define an event model for activities in monitoring videos.

February 9, 2016

- **Methods Using Supervised Classifiers** (Section 3) as long as the correct label of input data is known, supervised methods can be used effectively in learning process.
- **Methods Using Weakly-Supervised Classifiers** (Section 4) instead of supervised action recognition, it uses different unsupervised approaches to define long-term activity models. It also benifits from some supervised information as hints to help reach better models.
    - Using Visual Information and Text on Large Dataset
    - Using Visual Information and Audio
    - Using Visual Information, Text, Audio and etc.
- **Methods Using Unsupervised Classifiers** (Section 5) semi-supervised methods introduced to reduce the amount of required labeling tasks but in respect of long-term daily activities (ADLs) challenge is still there. The number of activities are large and they are quite different from each other and they are performed in variety of unconstrained environments. Unsupervised methods tend to develop generic models to recognize such activities.
- **Generating Complex Description: Verbalize, Machine Translation** (Section 6)
    - TACOS Cooking Dataset

2. Hand-Crafted, knowledge based activity recognition

Despite of significant advances within the field of Computer vision in the past decades, all mentioned approaches still rely on pixel- and feature-level data for the task, what tends to create a semantic gap between the model and the real-world event. Hand-crafted and logic-based models provide a explicit way to incorporate scene semantics.

Event detection methods in computer vision may be categorized in (adapted from Lavee et al. Lavee et al. (2009)): classification methods, probabilistic graphical models (PGM), and semantic models; which are themselves based on at least one of the following data abstraction level: pixel-based, feature-based, or event-based.

Artificial Neural Networks, Support-Vector Machines (SVM), and Independent Subspace Analysis (ISA) are examples of classification methods. For instance, Le et al.Le et al. (2011) have presented an extension of the ISA algorithm for event detection, where the algorithm learned invariant spatio-temporal features from unlabeled video data. Wang et al. Wang et al. (2011b) have introduced new descriptors for dense trajectory estimation as input for non-linear SVMs.

Common examples of PGMs approaches are Bayesian Network (BN), Conditional Random Fields, and Hidden Markov Models (HMM). BNs have been evaluated at the detection of person interactions (e.g., shaking hands) Park and Aggarwal (2004), left luggage Lv et al. (2006), and traffic monitoring Kumar et al. (2005). Nevertheless, this class of PGM has difficulty at modeling the temporal dynamics of an event. Izadinia and Shah Izadinia and Shah (2012) have proposed to detect complex events from by a graph representation of joint the relationship among elementary events and a discriminative model for complex event detection.

Even though the two previous classes of methods have considerably increased the performance of event detection in benchmark data sets, as they rely on pixel-based and feature-based abstractions they have limitations in incorporating the semantic and hierarchical nature of complex events. Semantic (or Description-based) models use descriptive language and logical operators to build event representations using domain expert knowledge. The hierarchical nature of these models allow the explicit incorporation of event and scene semantic with much less data than Classification and PGM methods. Zaidenberg et al. Zaidenberg et al. (2012) have evaluated a constraint-based ontology language for group behavior modeling and detection in airport, subways, and shopping center scenarios. Cao *et al.* Cao et al. (2009) have proposed an ontology for event context modeling associated to a rule-based engine for event detection in multimedia monitoring system. Similarly, Zouba *et al.* Zouba et al. (2010) have evaluated a video monitoring system at the identification of activities of daily living of older people using a hierarchical constraint-based approach.

Although Semantic models advantage at incorporating domain expert knowledge, the deterministic nature of their constraints makes them susceptible to noise from underlying components - *e.g.,* people detection and tracking components in a pipeline of computer vision system - as they lack a convenient mechanism to handle uncertainty. Probabilistic reasoning has been proposed to overcome these limitations. Ryoo and Aggarwal Ryoo and Aggarwal (2006) Ryoo and Aggarwal (2009a) have proposed hallucination concept to handle uncertainty from low-level components in a context-free grammar approach for complex event detection. Tran and Davis Tran and Davis (2008) have proposed Markov logic networks (MLNs) for event detection in parking lots. Kwak *et al.* Kwak et al. (2011) have proposed the detection of complex event by the combination of primitive events using constraint flows. Brendel et al Brendel et al. (2011) propose probabilistic event logic to extend an interval-based framework for event detection; by adopting a learned weight to penalize the violation of logic formulas. Similarly, Romdhane et al. Romdhane et al. (2013) proposed the use of weights to quantify the constraints utility for a constraint-based event detection.

3. Techniques Using Supervised Classifiers

3.1. Human Body Model Based Methods

Human body model based methods for action recognition use 2D or 3D information on human body parts, such as body part positions and movements. Typically, the pose of a human body is recovered and action recognition is based on pose estimation, human body parts, trajectories of joint positions, or landmark points.

Human body model based methods are inspired by a psychological research work of Johansson Jansson and Johansson (1973) on visual perception of motion patterns characteristics of living organisms in locomotion. Johansson has shown that humans can recognize actions from the motion of a few moving light displays attached to the human body, describing the motions of the main human body joints. He has found that between 10 and 12 moving light displays in adequate motion combinations in proximal stimulus evoke an impression of human walking, running, dancing, etc. (see Figure 1).

Yilmaz and Shah Yilma and Shah (2005) have proposed an approach for recognition of human actions in videos captured by uncalibrated moving cameras. The proposed approach is based on trajectories of human joint points.

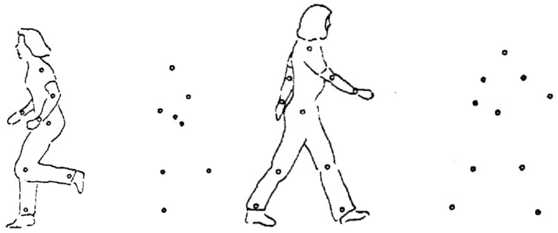


Figure 1: Johansson Jansson and Johansson (1973) Outline contours of a running and a walking subject, and the corresponding moving light displays attached to the human body.

In order to handle camera motion and different viewpoints of the same action in different environments, they use the multi-view geometry between two actions and they propose to extend the standard epipolar geometry to the geometry of dynamic scenes where the cameras are moving. Sample trajectories of the walking actions captured using a stationary camera and a moving camera are presented in Figure 2.

Ali et al. Ali et al. (2007) have also proposed an approach based on trajectories of reference joint points. These trajectories are used as the representation of the non-linear dynamical system that is generating the action, and they use them to reconstruct a phase space of appropriate dimension by employing a delay-embedding scheme. The properties of the phase space are captured in terms of dynamical and metric invariants that include Lyapunov exponent, correlation integral and correlation dimension. Finally, they represent an action by a feature vector which is a combination of these invariants over all the reference trajectories. Sample 3D trajectories generated by a head, two hands and two feet for the running action are presented in Figure 2.

Although all these techniques have shown to be promising, they have a big limitation. The extraction of human body model and body joint points in realistic and unconstrained videos is still a very difficult problem, and therefore these techniques remain limited in applicability.

The recent introduction of the cost-effective depth cameras helps in the extraction of human body joint points. The two most popular depth cameras are Microsoft Kinect and ASUS Xtion PRO LIVE motion sensor (see Figure 3). Both these sensors consist of an infrared pattern projector and an infrared camera to capture depth data, and a RGB camera to capture color images,

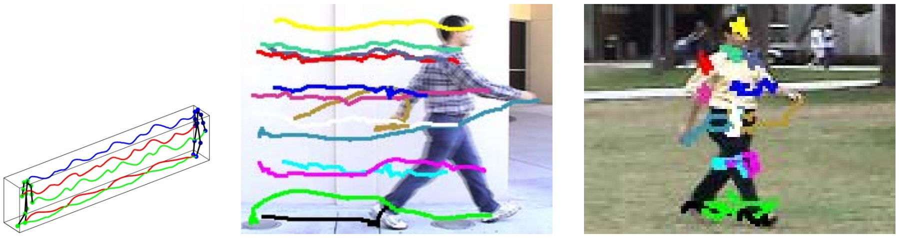


Figure 2: On the left: Ali et al. Ali et al. (2007): 3D trajectories generated by a head, two hands and two feet are shown for the running action. On the right: Yilmaz and Shah Yilma and Shah (2005): Trajectories of the walking actions are captured using a stationary camera and a moving camera.

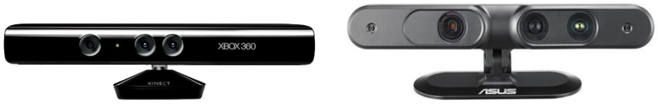


Figure 3: Microsoft Kinect on the left and ASUS Xtion PRO LIVE on the right.

see Figure 4. The depth cameras provide 3D depth data of the scene, which largely helps in people segmentation and in obtaining the 3D joint positions of the human skeleton. Several techniques that use such depth cameras and the extracted human skeleton have been proposed, e.g. Raptis et al. Raptis et al. (2011) and Wang et al. Wang et al. (2012). However, the cost-effective depth cameras also have some limitations.

- First of all, the range of the depth sensor is limited, e.g. Microsoft recommends to use the Kinect sensor in the range between 0.5 m and 4 m 1, and ASUS recommends to use the Xtion PRO LIVE motion sensor in the range between 0.8 m and 3.5 m 2. Although it is possible to use the depth data at larger distances, the quality of the data is degraded by the noise and low resolution of the depth measurements. For example, it is possible to get the depth data even up to 10 meters from the Microsoft Kinect Litomisky (2012), but Khoshelham and Elberink (Khoshelham and Elberink, 2012) show that: (a) the random error of depth data increases quadratically with increasing distance and reaches 4 cm at the range of 5 meters, (b) the depth resolution decreases quadratically with increasing distance and the point spacing in the depth direction reaches 7 cm at the range of 5 meters, and (c) for indoor mapping applications the data should be acquired within 1-3 m distance to the sensor. Human pose estimation in such motion sensors is typically extracted from depth images, e.g. Shotton et al. Shotton et al. (2013) (see Figure 4), and as a result the quality of human pose estimation algorithms decreases with increasing distance.

- Second of all, skeleton tracking and the estimated 3D joint positions are noisy and can produce inaccurate results or even fails when serious occlusion occurs [Wang 2012], e.g. when one leg is in front of the other, a hand is touching another body part, or two hands are crossing.

Therefore, we focus on action recognition using RGB cameras due to many potential applications of such sensors.

3.2. Holistic Methods

Shape and silhouette information based features are one of the very first characteristics, which were used to represent human body structure and its dynamics for action recognition in videos.

One of the first approaches using silhouette images and features for action recognition is the work of Yamato et al. Yamato et al. (1992). They extract a human shape mask for each image, calculate a grid over the silhouette, and for each cell of the grid calculate the ratio of foreground to background pixels (see Figure 5). Then, each grid representation of an image is assigned to a symbol, which corresponds to a codeword in the codebook created by the Vector Quantization technique. Finally, Hidden Markov Models (HMMs) are applied for action recognition and the model which best matches the observed symbol sequence is chosen as the recognized action category.

Bobick and Davis Bobick and Davis (2001) were first to introduce the idea of temporal templates for action recognition. They extract human shape masks from images and accumulate their differences between consecutive frames. These differences are then used to construct a binary motion-energy image (MEI) and a scalar-valued motion-history image (MHI) (see Figure 6). The former indicates the presence of motion, and the latter represents the recency of motion, i.e. the pixel intensity is a function of the temporal history of motion at that point. Then, they proposed a recognition method matching temporal templates against stored instances of actions. The MEI and MHI together can be considered as a two component version of a temporal template.

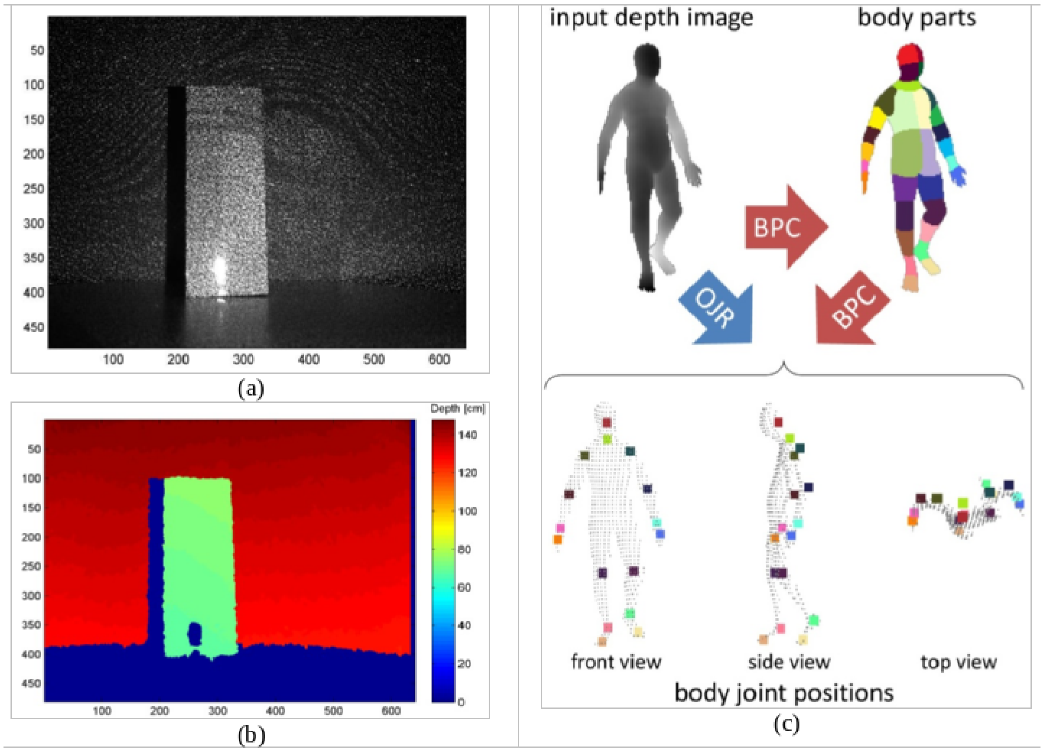


Figure 4: (a) Sample infrared image, from the Microsoft Kinect, presenting pattern of speckles projected on a sample scene Litomisky (2012). (b) The resulting depth image (Litomisky, 2012). (c) Two sample approaches for estimating human pose from single depth images Shotton et al. (2013). Body part classification (BPC) predicts a body part label at each pixel (labels are represented by colors), and then uses these labels to localize the body joints. Offset joint regression (OJR) more directly regresses the positions of the joints.

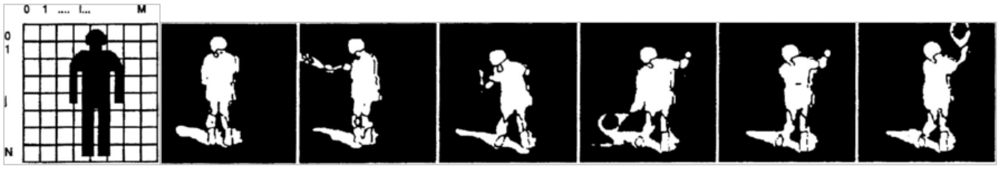


Figure 5: Yamato et al. Yamato et al. (1992) Mesh feature (the first image), and the sample shape masks for the forehand stroke action from the tennis action (the remaining images).

Blank et al. Blank et al. (2005) proposed a model based on three-

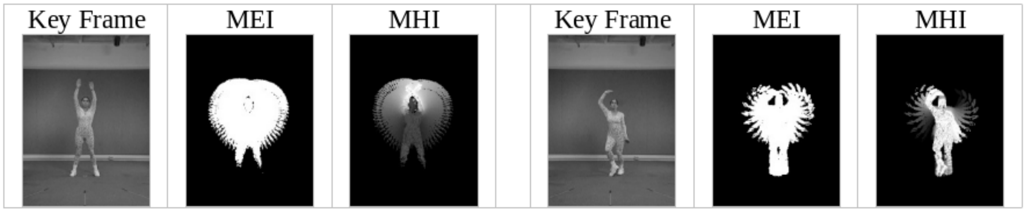


Figure 6: Bobick and Davis Bobick and Davis (2001) MEI and MHI representations calculated for two sample actions (move 4 and move 17) together with sample key frames.

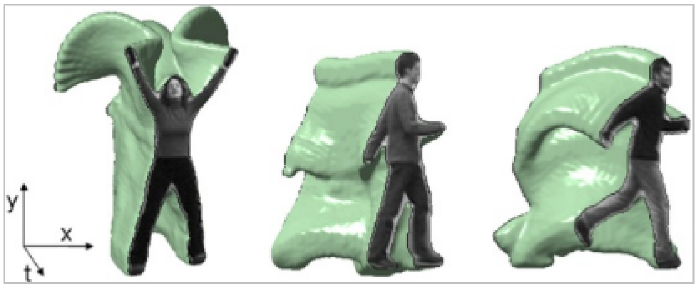


Figure 7: Blank et al. Blank et al. (2005) Sample spatio-temporal volumes constructed by stacking silhouettes over a given sequence.

dimensional shapes induced by the silhouettes in the space-time volume. At each frame, they compute a silhouette information using a background subtraction technique. They stack silhouettes over a given sequence to form a spatio-temporal volume (see Figure 7). Then, they use 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. They use chunks of 10 frames length and match these chunks using a sliding window approach. The action classification is done using simple nearest neighbour algorithm with an Euclidean distance. The main disadvantage of the holistic based method is the requirement of shape, silhouette extraction, what is typically done by segmentation. The accuracy of these techniques is highly related to the correctness of the segmentation and the precise segmentation is very difficult to obtain in real world videos.

3.3. Local Feature Methods

Action recognition based on local features is one of the most active research topics. The main advantage of the local features based methods is that no information on human body model or localization of people is required. In this section, we focus on local feature methods.

3.3.1. Local Features

Local features are extracted by applying a local feature detector and then by encoding spatio-temporal neighbourhoods around the detected features using a local feature descriptor. In this section we describe the most popular local spatio-temporal detectors (see Section 3.3.2) and descriptors (see Section 3.3.3) for action recognition in videos.

3.3.2. Local Feature Detectors

Local feature detectors for videos can be divided into two categories: spatio-temporal interest point detectors and trajectory detectors.

3.3.2.1. Spatio-Temporal Interest Point Detector. One of the first works on local feature detectors for videos is the work of Laptev and Lindeberg (Laptev, 2005), see Figure 8. They proposed the Harris3D interest point detector, which is an extension of the Harris detector (Harris and Stephens, 1988) to the spatio-temporal domain by requiring the video values in space-time to have large variations in both the spatial and the temporal dimensions. The Harris3D detector calculates a spatio-temporal second-moment matrix at each video point and searches for regions that have significant eigenvalues of the matrix. The final spatio-temporal points are detected as local positive spatio-temporal maxima. Moreover, the detected points have to be the local extrema of the normalized spatio-temporal Laplace operator, which is defined to select the spatio-temporal scales of points.

Dollar et al. Dollár et al. (2005) observed that sometimes true spatio-temporal corners are rare, even when interesting motion occurs, and might be too rare in certain cases, e.g. for face expression recognition. Therefore, they proposed the Gabor detector, which gives denser results than the Harris3D. The Gabor detector applies a set of spatial Gaussian kernels and temporal Gabor filters. The final spatio-temporal points are detected as local maxima of the defined response function.

Different from the above, Oikonomopoulos et al. Oikonomopoulos et al. (2005) proposed a space-time extension of a salient region detector Kadir and Brady (2003) using entropy. The proposed detector selects the scales at

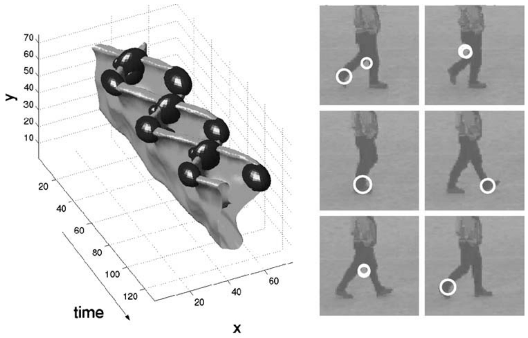


Figure 8: Laptev (Laptev, 2005) the Harris3D interest points from the motion of the legs of a walking person; left image: 3D plot with a thresholded level surface of a leg pattern (upside down) and the detected points (ellipsoids); right image: interest points overlayed on single frames in the original video sequence.

which the entropy achieves local maxima and forms spatio-temporal salient regions by clustering spatio-temporal points with similar location and scale.

Willems et al. Willems et al. (2008) proposed the Hessian3D interest point detector, which is a spatio-temporal extension of the Hessian saliency measure for blob detection in images Beaudet (1978). The Hessian3D detector calculates the Hessian matrix at each interest point and uses the determinant of the Hessian matrix for point localization and scale selection. The detector uses integral video to speed up computations by approximating derivatives with box-filter operations. The detected points are scale-invariant and dense, typically they are denser than from the Harris3D detector but not that dense as from the Gabor detector.

Most of the techniques use local information to detect spatio-temporal interest points. Wong and Cipolla Wong and Cipolla (2007) proposed an interest point detector which uses global information, i.e. the organization of pixels in a whole video sequence, by applying non-negative matrix factorization on the entire video sequence. The proposed detector is based on the extraction of dynamic textures, which are used to synthesize motion and identify important regions in motion. The detector extracts structural information, e.g. the location of moving parts in a video, and searches for regions that have a large probability of containing the relevant motion.

Different from the above techniques, Wang et al. Wang et al. (2009) proposed to apply dense sampling. The dense sampling extracts interest points at regular positions and scales in space and time. The sampling is done using 5 dimensions (x, y, t,σ ,τ, where (x, y, t) is the spatio-temporal position of a point, σ is the spatial scale, and τ is the temporal scale. This detector extracts a big amount of features but is also able to extract relevant video features.

When faced with the decision Which Spatio-Temporal Interest Point detector gives the best results?, there is no clear answer. Wang et al. Wang et al. (2009) compared Harris3D, Gabor detector, Hessian3D, and dense sampling. The comparison was done on three datasets: (a) KTH dataset, where Harris3D achieved the best results, (b) UCF dataset, where dense sampling achieved the best results, and (c) Hollywood2 dataset, where dense sampling achieved the best results using reference videos, but Harris3D with full resolution videos achieved better results than the dense sampling with reference videos. Therefore, according to that evaluation, there is no single detector that always achieves the best results, but among the four selected detectors (i.e. Harris3D, Gabor detector, Hessian3D, and dense sampling), the best results per dataset are achieved either by Harris3D or dense sampling.

3.3.2.2. Trajectory Detector. Trajectories are typically extracted by detecting interest points and tracking them in the consecutive frames.

One of the best-known feature tracking algorithm is the KLT (Kanade-Lucas-Tomasi) [Lucas et al. (1981), Tomasi and Kanade (1991), Shi and Tomasi (1994)]. The KLT algorithm locates good features for tracking by examining the minimum eigenvalue of each 2X2 gradient matrix, and then features are tracked using a Newton-Raphson method of minimizing the difference between the two windows. Sample work using the KLT tracker is the work of Matikainen et al. Matikainen et al. (2009), see Figure 9, where they extract trajectories of fixed length using a standard KLT tracker and then cluster the trajectories. They compute an affine transformation matrix for each cluster center, and the elements of the matrix are then used to represent the trajectories. Messing et al. Messing et al. (2009) proposed to apply a different detector of points, Harris3D detector, and track points with the KLT tracker. Then, the trajectories, which vary in length, are represented as sequences of log-polar quantized velocities and used for action classification.

Kaaniche and Bremond Kaaniche and Brémond (2009) proposed to detect interest points using Shi and Thomasi corner detector Shi and Tomasi

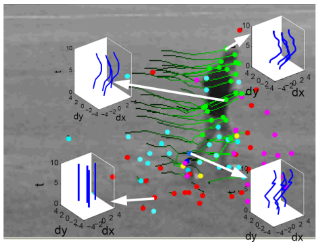


Figure 9: Matikainen et al. Matikainen et al. (2009): Feature points are tracked using the KLT tracking algorithm, and then the trajectories are clustered and assigned to the library of trajectories.

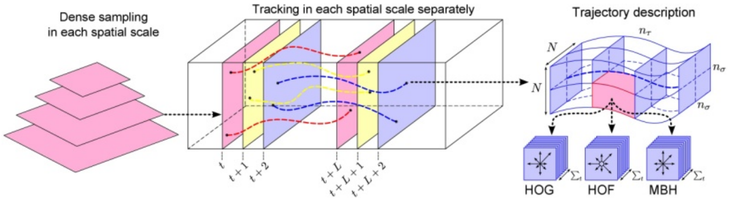


Figure 10: Wang et al. Wang et al. (2010) Overview of the dense trajectories; left image: dense sampling on multiple spatial scales; middle image: feature tracking in the corresponding spatial scale over L frames; right image: descriptors calculated around a trajectory.

(1994) or Features from Accelerated Segment Test (FAST) corner detector Rosten and Drummond (2006), and then track points using matching the HOG descriptors over consecutive frames. The obtained trajectories vary in length and according to the authors are less sensitive to the noise than the trajectories from the KLT tracker.

Different from the above techniques, Sun et al. Sun et al. (2009) proposed to extract trajectories based on the pairwise SIFT matching over consecutive frames. They claim that scale-invariant properties of the SIFT descriptor is a better choice when compared to the Harris and KLT based feature trackers.

In Sun et al. (2010), Sun et al. proposed to combine the tracking results of the KLT and the SIFT trackers, and formulated the visual matching and tracking in a unified constrained optimization problem. In order to extract dense trajectories, the authors add interior points that are neither corner points tracked by the KLT nor by the SIFT trackers by interpolating of the surrounding flows, subject to block-matching constraints. Wang et al. Wang et al. (2010) also proposed to extract dense trajectories, see Figure 10. They apply dense sampling to extract interest points and track them using a dense optical flow field. Then, the trajectories are represented using the trajectory shape, HOG, HOF and MBH descriptors. This technique gives a great number of trajectories but according to the authors they obtain better results than the trajectories from the KLT and the SIFT tracking algorithms. When faced with the decision Which trajectory detector gives the best results?, we refer to the work of Wang et al. Wang et al. (2013), where:

- the dense trajectories were compared with trajectories extracted by Kanade-Lucas-Tomasi (KLT) tracker and by SIFT descriptor matching. In all cases, i.e. on nine datasets, the dense trajectories outperformed the other trajectories.

- the dense trajectories were compared with the trajectories from Sun et al. (2010) and from Messing et al. (2009) on the KTH dataset, and the dense trajectories outperformed the other trajectories.

3.3.3. Local Feature Descriptors

Local feature descriptors capture shape and motion information in a local neighborhood surrounding interest points and trajectories.

One of the first works on local feature descriptors for videos is the work of Laptev and Lindeberg Laptev and Lindeberg (2006). They presented and compared several descriptors based on motion representations in terms of spatio-temporal jets (higher-order derivatives), position dependent histograms, position independent histograms, and principal component analysis computed for either spatio-temporal gradients or optical flow. They reported the best results for descriptors based on histogram of spatio-temporal gradients and optical flow.

Dollar et al. Dollár et al. (2005) also proposed several local feature descriptors. They considered three transformations to local neighborhoods: normalized pixel values, the brightness gradient, and windowed optical flow. They also considered three methods to create a feature vector: flattening the local neighborhood into a vector, histogramming the values in the local neighborhood, and dividing the local neighborhood into a grid and histogramming the values in each cell of the grid. For all methods, the PCA was applied to reduce the dimensionality of the final descriptors. They reported the best results for descriptors based on concatenated gradient information.

The HOG (Histogram of Oriented Gradients) and HOF (Histogram of Optical Flow) are the popular local feature descriptors for videos proposed by Laptev et al. Laptev et al. (2008). The HOG descriptor for videos is the variant of the HOG image descriptor Dalal and Triggs (2005). In order to embed structure information in a descriptor, the local neighborhood surrounding a local feature is divided into a spatio-temporal grid. For each cell of the grid, a histogram descriptor is calculated. Then, the histograms are normalized and concatenated into the final descriptor. The HOG descriptor encodes visual appearance and shape information; the edge orientations are calculated and quantized into histogram bins. The HOF descriptor encodes motion information; the optical flow is calculated and quantized into histogram bins.

The 3DSIFT (3-Dimensional SIFT) is an extension of the SIFT (Scale Invariant Feature Transform) image descriptor Lowe (2004) to the spatio-temporal domain proposed by Scovanner et al. Scovanner et al. (2007). It is based on the spatio-temporal grid idea and spatio-temporal gradients. Each pixel is weighted by a Gaussian centered on the given position and votes into a grid of histograms of oriented gradients. A Gaussian weighting is applied to give less importance to gradients farther away from the local feature center. To be rotation-invariant, a dominant orientation is determined and is used for orienting the grid descriptor.

The HOG3D descriptor is another extension of the HOG image descriptor Dalal and Triggs (2005) to the spatio-temporal domain proposed by Klaser et al. Klaser et al. (2008). The HOG3D is based on the spatio temporal grid idea and 3D gradients, which are calculated and quantized to the histograms of 3D gradient orientations based on convex regular polyhedrons.

The main differences between the HOG, ESIFT, and the HOG3D spatio-temporal descriptors are: (1) the HOG descriptor only considers spatial gradients, and the ESIFT and the HOG3D descriptors consider spatio-temporal 3D gradient orientation, and (2) the ESIFT descriptor uses regular binning based on spherical coordinates, and the HOG3D descriptor uses regular polyhedrons and spherical coordinates for which the amount of bins can be controlled separately for spatial and temporal gradient orientations.

The ESURF (Extended SURF) is an extension of the SURF (Speeded Up

Robust Features) image descriptor Bay et al. (2006) to the spatio-temporal domain proposed by Willems et al. Willems et al. (2008). The ESURF divides the local neighborhood surrounding a local feature into a spatio-temporal grid, and it represents each cell of the grid by a vector of weighted sums of uniformly sampled responses of Haar-wavelets along the three (x, y, z) axes.

The MBH (Motion Boundary Histogram) is an extension of the MBH image descriptor Dalal et al. (2006) to the spatio-temporal domain proposed by Wang et al. Wang et al. (2011a). The MBH descriptor separates the optical flow field into its x and y components. Spatial derivatives are computed separately for the horizontal and vertical components of the optical flow, and orientation information is quantized into histograms, similarly to the HOG descriptor. The MBH descriptor is also based on the spatio-temporal grid idea.

The Trajectory shape descriptor was proposed by Wang et al. Wang et al. (2011a) to encode a shape of the extracted dense trajectories. It describes a shape of a trajectory by a sequence of displacement vectors normalized by the sum of displacement vector magnitudes.

When faced with the decision Which local feature descriptor should we use?, we refer to the work of:

- Wang et al. (2009), which recommends using the combination of HOG and HOF descriptors for the Spatio-Temporal Interest Points.
- Wang et al. (2013), which recommends using the combination of Trajectory shape descriptor, HOG, HOF, and MBH descriptors for the Dense Trajectories. This combination achieved the best results on 8 out of 9 datasets, when compared with each of the descriptors separately; the best result for the remaining dataset was achieved by the MBH descriptor alone. The authors underline the importance of the MBH descriptor, which is robust to camera motion.

3.3.4. Collections of Local Features

The methods based on local features presented in the previous section (Section 4.1) are based on the discriminative power of individual local features and global statistics of individual local features. Although these techniques have shown very good results in action recognition, they also have a few limitations:

- they ignore position of features,
- they ignore local density of features,
- they ignore relations among the features (i.e. visual appearance and motion relations, spatio-temporal order among features, and spatio-temporal geometric relations among features (i.e. Δx,Δy,Δt)).

These techniques might distinguish various actions but may fail to distinguish similar actions as they do not use all the available information. A common way to overcome these limitations is to use either spatio-temporal grids Laptev et al. (2008) or multi-scale pyramids Lazebnik et al. (2006). However, these techniques are still limited in terms of detailed description providing only a coarse representation.

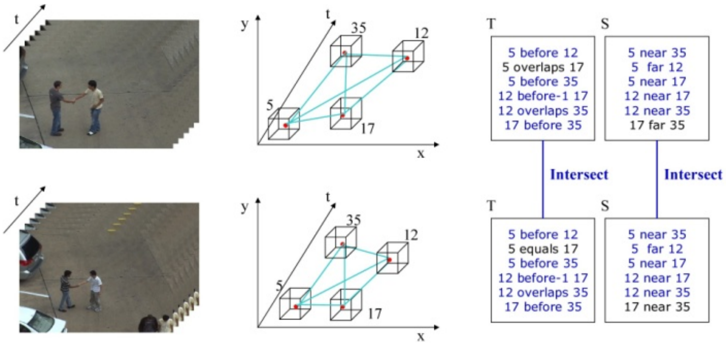
In order to cope with these problems, several solutions have been proposed, most of which try to create higher-level feature representations and use them together with the bag-of-features approach. These higher-level feature representations we can divide into 2 categories:

- **Pairwise Features**- features capturing pairwise relations among features.
- **Contextual Features** - features capturing relations among any number of neighboring features.

These higher-level feature representations have shown to enhance the discriminative power of individual local features and improve action recognition accuracy.

3.3.5. Pairwise Features

One of the first studies on pairwise features is the work of Liu et al. Liu and Shah (2008). They proposed to explore the correlation of the compact visual word clusters using a modified correlogram. Firstly, they extract local features using the detector and the descriptor proposed by the Dollar et al. Dollár et al. (2005). Then, they represent a video sequence using



(a) 3-D XYT Video volumes (b) Spatio-temporal feature points (c) Relationships in histograms

Figure 11: Ryoo et al. Ryoo and Aggarwal (2009b) Spatio-temporal relationship matching process: (a) two given videos, (b) extraction of local features and calculation of pairwise relations, (c) calculation of a relationship histogram per input video, and similarity between relationship histograms calculated as intersection.

the bag-of-features approach. Instead of using the k-means algorithm, they apply Maximization of Mutual Information to discover the optimal number of codewords. Then, to capture the structural information they explore the correlation of the codewords. They apply the modified correlogram, which is somewhat scale invariant, translation and rotation invariant. As they calculate the probability of co-occurence between every pair of codewords, they use small codebooks. Ryoo et al. Ryoo and Aggarwal (2009b) proposed a spatio-temporal relationship matching technique, which is designed to measure structural similarity between sets of features extracted from two videos (see Figure 11). Firstly, the authors extract local features for every video sequence. Then, they create pairwise relations among features, and represent each video sequence using relationship histograms. The relationship histogram is created separately both for the spatial and the temporal order, and it is based on simple, constant and limited predicates indicating the order of features. Then, the authors compute the relationship histograms intersection to measure similarity between two videos. The main limitations of this technique are: (a) the relationship histograms use only simple predicates (e.g. before and after) to encode pairwise relations between local features, (b) the spatial and the temporal orders between local features are encoded independently and not both at the same time, and (c) the spatio-temporal geometric relations (i.e. Δx, Δy, Δt) among features are ignored. Ta et al. Ta et al. (2010) proposed pairwise features, which encode both appear-

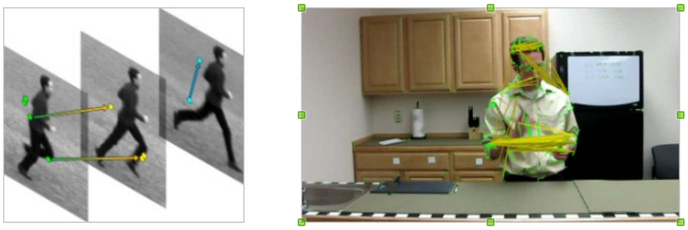


Figure 12: On the left: Ta et al. Ta et al. (2010) Sample pairwise features are presented as local features Dollár et al. (2005) detected as close in time and close in space. On the right: Matikainen et al. Matikainen et al. (2010) Sample pairwise features are presented as pairs of local features selected to be discriminative for a specific action class.

ance and spatio-temporal relations of local features. Firstly, the authors extract the Spatio-Temporal Interest Points (STIPs) from a video sequence. Then, the pairwise features are created by grouping pairs of STIPs, which are both close in space and close in time. The pairwise features are encoded by appearance and spatio-temporal relations of local features. The appearance relations are captured by concatenating the appearance descriptors of STIPs. The spatio-temporal relations are captured by a spatio-temporal distance between STIPs. Then, for each type of relations the bag-of-features approach is applied independently and the two obtained representations are concatenated. The main limitations of this technique are: (a) it is difficult to correctly set the spatial and temporal thresholds to decide which STIPs are both close in space and close in time, (b) spatio-temporal order between features is lost, and (c) association between appearance and the spatio-temporal geometric information is lost by calculating two independent codebooks.

Matikainen et al. Matikainen et al. (2010) also proposed a method for representing spatiotemporal relationships between features in the bag-of-features approach (see Figure 12). The authors use both the Spatio Temporal Interest Points (STIPs) and trajectories to extract local features from a video sequence. Then, they combine the power of discriminative representations with key aspects of Naive Bayes. As the number of all possible pairs and relationships between features is big, they reduce the number of relationships to the size of the codebook. Moreover, they show that the combination of both the appearance and motion base features improves the action recognition accuracy. The main limitation of this technique is that it encodes the appearance and motion relations among features but it does not use information about the spatio-temporal geometric relations between features.

Banerjee et al. Banerjee and Nevatia (2011) proposed to model pairwise co-occurrence statistics of visual worlds. Firstly, the authors extract local features and they create a codebook of local features represented by local descriptors. Instead of selecting the most discriminative relations between features, they use small codebooks, i.e. the codebook size is smaller than 20. They model local neighborhood relationships between local features in terms of a count function which measures the pairwise co-occurrence frequency of codewords. Then, the count function is transformed to the edges connecting the latent variables of a Conditional Random Field classifier, and they explicitly learn the co-occurrence statistics as a part of its maximum likelihood objective function. The main limitations of this technique are: (a) it can only use small codebooks, and (b) it uses discriminative power of individual (appearance) features but information about the spatio-temporal geometric relations and spatio-temporal order between features is ignored.

In summary, most of the above pairwise features based techniques use the discriminative power of individual features and capture visual relations among features. However, the existing techniques ignore information about spatio-temporal geometric relations between features (i.e. Δx, Δy, Δt) and spatio temporal order between features. Moreover, some of the above techniques can only handle small codebooks [Liu and Shah (2008), Banerjee and Nevatia (2011)] due to quadratic processing time. Therefore, a new and optimized representation is needed to create a finer description of pairwise features.

3.3.6. Contextual Features

Pairwise features only capture relations between two features. Contextual features are able to capture relations among many features. One of the first studies on contextual features is the work of Sun et al. Sun et al. (2009) (see Figure 13). They proposed to model the spatio-temporal context information of video sequences based on the SIFT based trajectories. The spatio-temporal context is represented in a hierarchical way: point-level, intra-trajectory, and inter-trajectory context. The point-level context is measured as the average of all the SIFT features extracted around the trajectory. The intra-trajectory context is encoded as the transition and dynamics of the trajectory in spatio-temporal domain. The inter-trajectory context (see Figure 13) is represented as contextual features and captures local occurrence statistics of quantized

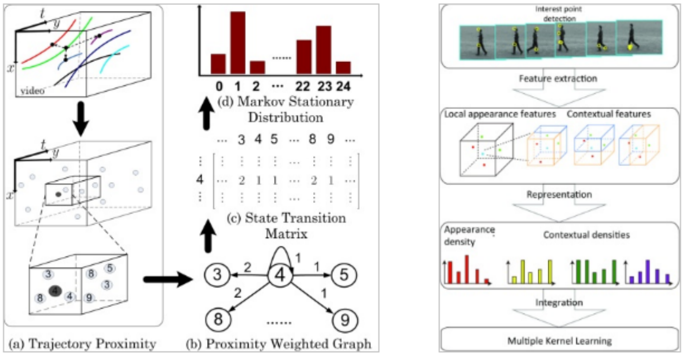
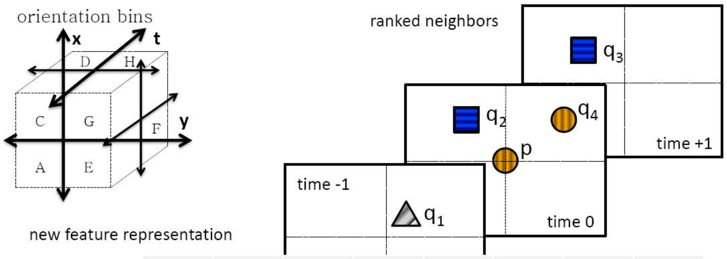


Figure 13: Sun et al. Sun et al. (2009) The proposed inter-trajectory context representation (left image). Wang et al. Wang et al. (2011c) Overview of the proposed approach (right image).

trajectories within figure-centric neighbourhoods. The intra-trajectory and inter-trajectory context encode the spatio-temporal context information into the transition matrix of a Markov process, and extract its stationary distribution as the final context descriptor. The main limitations of the proposed contextual features are: (a) they ignore pairwise relations among features, and (b) they ignore spatio-temporal geometric relations among features.

Similarly, Wang et al. Wang et al. (2011c) proposed to capture contextual statistics among interest points based on the density of features observed in each interest points contextual domain (see Figure 13). Firstly, the authors extract local features for a given video sequence. Then, they create spatio-temporal contextual features that capture contextual interactions between interest points, i.e. they capture the density of all features observed in each interest points mutliscale spatio-temporal contextual domain. Then, they apply the bag-of-features approach for local features and contextual features, and augment the obtained video representations using Multiple Kernel Learning approach. The main limitations of the proposed contextual features are: (a) they ignore pairwise relations among features, and (b) they ignore spatio temporal geometric relations and spatio-temporal order among features. Kovashka et al. Kovashka and Grauman (2010) proposed figure-centric statistics that capture the orientation among features, see Figure 14. Firstly, the authors extract local features from videos, i.e. they either apply: (1) dense sampling and represent interest points using HOG3D de-



	bin A			bin B			bin C			bin D			bin E			bin F			bin G			bin H		
counts for	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲	●	■	▲
neighbors 0 to 0																						1		
neighbors 0 to 1																					1	1		
neighbors 0 to 2											1										1	1		
neighbors 0 to 3											2										1	1		
neighbors 0 to 4											2										1	2		

Figure 14: Kovashka et al. Kovashka and Grauman (2010) Contextual Features. A figure-centric neighbourhood divided into 8 orientations, three frames with sample features, and the histogram representation of the neighbourhood.

scriptors, or (2) Harris3D detector and represent interest points using HOG and HOF descriptors. Then, they create a visual vocabulary, they quantize localfeatures, and use the quantized features to create figure-centric features consisting of the words associated with nearby points and their orientation with respect to the central interest point. Such figure-centric features are then recursively mapped to higher-level vocabularies, producing a hierarchy of figure-centric features. Moreover, the authors propose to learn the shapes of space-time feature neighborhoods that are the most discriminative for a given action category. The main limitations of this technique are: (a) only the orientation among feature points is captured and not the spatio-temporal distance relations among features, and (b) the contextual features are of high dimension (40Xk, where k is the codebook size) and the clustering process of these contextual features might be time consuming.

All techniques presented above use the bag-of-features approach in order to encode the created contextual features. The techniques presented below create contextual features but do not use the bag-of-features approach.

Gilbert et al. Gilbert et al. (2009) proposed to use dense corner features that are spatially and temporally grouped in a hierarchical process. They build compound hierarchical features, which can be seen as contextual features, based on relationships of detected interest points, and they find frequently reoccurring patterns of features using data mining.

The local features are represented only using scale, channel, and dominant orientation of features. The main limitations of this technique are: (a) it does not use visual and motion appearance information, and (b) it ignores pairwise relations among features and information about spatio-temporal order among features.

Oshin et al. Oshin et al. (2011) proposed another contextual features and they use the spatio-temporal distribution of features alone, i.e. without explicit appearance information. Their approach makes use of locations and strengths of interest points only, and it discards appearance information. In order to automatically discover and reject outlier samples within classes, they use Random Sampling Consensus (RANSAC). The main limitations of this technique are: (a) it does not use visual and motion appearance information, and (b) it ignores pairwise relations among features and information about spatio-temporal order among features.

In summary, most of the above contextual features based techniques use the discriminative power of individual features and capture local density of features in feature-centric neighbourhoods. To capture structural information in contextual features, the spatiotemporal grid has been applied in some of the above approaches, however the spatiotemporal grid is limited in terms of detailed description providing only a coarse representation.

Moreover, the existing techniques ignore information about the spatio-temporal order among features. Therefore, a new representation is needed to create a finer description of contextual features.

3.3.7. Local Features Encoding

Once local features are extracted, they are used to represent videos - actions. The most popular representation technique encoding local features is the bag-offeatures model. The bag-of-features is a very popular representation used in Natural Language Processing, Information Retrieval, and also Computer Vision. It was originally proposed for document retrieval, where text is represented as the bag of its words (bag-of-words) Salton (1968).

One of the first and important studies using bag-of-features model in Computer Vision are: Cula and Dana Cula and Dana (2001) for texture classification, Sivic and Zisserman Sivic and Zisserman (2003) for object and scene retrieval, Csurka et al. Csurka et al. (2004) for image categorization, Lazebnik et al. Lazebnik et al. (2006) for scene categorization, Sivic et al. Sivic et al. (2005) for object localization, and Schuldt et al. Schüldt et al. (2004), Dollar et al. Dollár et al. (2005), and Niebles et al. Niebles et al. (2008) for action recognition.

The bag-of-features model encodes global statistics of local features, computing a spatial histogram of local feature occurrences in a video sequence. Firstly, it creates a visual vocabulary using unsupervised learning over local features extracted from the training videos. The learning is typically done with k means clustering algorithm. Then, the bag-of-features quantizes local features to a visual vocabulary, and it represents a video using histogram of quantized local features, followed by the L1 or the L2 norm; both norms are popular and there is no clear answer which one is the best. The advantage of the L1 norm is that it requires less computation time. The normalization step is applied to reduce effects of variable video size and variable number of detected local features in videos.

The bag-of-features model uses hard quantization of local features (i.e. uses histogram encoding) to represent local features. Recent approaches replace the hard quantization of local features with alternative encoding techniques that retain more information about the local features. This has been done in two ways: (1) by representing features as a combination of visual words (e.g. Kernel codebook encoding [Philbin et al. (2008), van Gemert et al. (2008)] and Locality-constrained Linear Coding Wang et al. (2010)), and (2) by representing differences between features and visual words (e.g. Fisher vector encoding Perronnin et al. (2010), Super-vector encoding Zhou et al. (2010), BossaNova encoding Avila et al. (2013), and Vector of Locally Aggregated Descriptors encoding Jégou et al. (2010)). A good description of various encoding techniques is provided in Chatfield et al. (2011), where the encoding techniques are applied for object recognition (but can be applied for action recognition as well).

The following techniques are based on visual vocabulary, which is typically created in the same manner as in the bag-of-features model, unless otherwise stated.

Kernel codebook encoding [Philbin et al. (2008), van Gemert et al. (2008)] is a variant of the bag-of-features model, where local features are assigned to visual vocabulary in a soft manner. The local features are associated with several nearby visual words instead of a single nearest visual word, and they are mapped to a weighted combination of visual words.

Locality-constrained Linear Coding [Wang et al. (2010), Zhou et al. (2013)] is another variant of the bag-of-features approach. It projects each local feature into its local-coordinate system, and the projected coordinates are integrated by max pooling technique to generate the final representation. Features are projected down to the local linear subspace spanned by several closest visual words.

Fisher vector encoding (Fisher vectors) [Perronnin et al. (2010), Oneata et al. (2013)] does not represent features as a combination of visual words but instead it represents differences between features and visual words. Firstly, it creates a visual vocabulary by clustering local features extracted from the training videos, where clustering is done with Gaussian Mixture Model clustering. Then, it captures the average first and second order differences between local features and visual vocabulary, i.e. Gaussian components.

Super-vector encoding Zhou et al. (2010) is another variant of the Fisher encoding. There are two variants of the support vector encoding: (1) with hard assignment of local features to the nearest visual word, and (2) with soft assignment of local features to several nearest visual words. The visual vocabulary is created using k-means algorithm. Then, the video is encoded using (1) the first order differences between local features and visual words and (2) the components representing the mass of each visual word. Vector of Locally Aggregated Descriptors (VLAD) encoding Jégou et al. (2010), Jain et al. (2013) is another variant of the bag of-features model. It accumulates the residual of each local feature with respect to its assigned visual word. Then, it matches each local feature to its closest visual word. Finally, for each cluster it stores the sum of the differences of the descriptors assigned to the cluster and the centroid of the cluster.

BossaNova encoding Avila et al. (2013) is very similar to the Vector of Locally

Aggregated Descriptors encoding technique. It enriches the bag-of-features representation with a histogram of distances between the local features and visual words, preserving information about the distribution of the local feature around each visual word.

Most of the above techniques were invented for image classification, image retrieval, and object recognition. However, they can be applied for any domain and any task using local features.

Features Encoding: Memory Requirements Lets denote the size of codebook as K and the size of local descriptors as D. Then:

- The size of the bag-of-features representation, Kernel codebook encoding, and Locality-constrained Linear Coding is K.
- The size of the Fisher vector encoding is 2KD.
- The size of the VLAD encoding is KD.
- The size of the BossaNova encoding is K(B + 1), where B is the number of discretized distances between codewords and local descriptors Avila et al. (2013).

The bag-of-features, Kernel codebook encoding, and Localityconstrained Linear Coding representations require the smallest amount of memory to store a video sequence. The Fisher vector encoding requires the greatest amount of memory to store a video sequence.

Local Features Encoding: Accuracy Various comparisons between local feature encoding techniques have been presented in the literature, e.g.:

- Chatfield et al. Chatfield et al. (2011) compared the bag-of-features, Kernel codebook encoding, Locality-constrained Linear Coding, Fisher vector encoding, and Supervector encoding, and for the task of object recognition Fisher vector encoding gave the best results.
- Avila et al. Avila et al. (2013) compared the bag-of-features, BOSSA encoding Avila et al. (2011), BossaNova encoding (improved version of the BOSSA encoding), and Fisher vector encoding, and for the task of image classification Fisher vector encoding gave the best results (not

counting the combination of the Fisher vector encoding and BossaNova encoding which shown superior results).

- Moreover, Jegou et al. Jégou et al. (2012) compared the bag-of-features, Fisher vector encoding and VLAD encoding, and for large-scale image search again Fisher vector encoding gave the best results.
- Krapac et al. Krapac et al. (2011) compared the bag-of-features and the Fisher vector encoding, and for image categorization again Fisher vector encoding gave the best results.
- For large-scale web video event classification, Sun and Nevatia Sun and Nevatia (2013) presented that the Fisher vector encoding obtained better results than the bag-of-features and the VLAD encoding.
- Similarly, for the action recognition task, Oneata et al. Oneata et al. (2013) presented that the Fisher vector encoding obtained better results than the bag-of-features representation.

Local Features Encoding: Conclusion The bag-of-features approach is the most popular technique for encoding local features and its representation requires a small amount of memory to store a video sequence. The recent Fisher vector encoding seems to be very powerful technique, it has shown superior results for many Computer Vision tasks, but its representation requires a large amount of memory to store a video sequence. Fortunately, it has been shown Perronnin et al. (2010) that the Fisher vector encoding can be used with linear classifiers and it still outperforms the bag-of-features representation, which should be applied with non-linear classifiers to give a good classification performance.

3.4. Classifiers

Once we represent video sequences, e.g. using any of the above techniques, we would like to decide which actions they contain. We are given a set of actions and our goal is to recognize these actions in videos. Due to a large number of machine learning algorithms, we only briefly present several popular classification algorithms.

The goal of the supervised learning is to build a model of the distribution of class labels in terms of input features. Then, the obtained classifier assigns class labels to the testing instances, where the values of the input features are known, but the value of the class label is unknown.

Statistical approaches Jensen (1996) provide a probability that a given instance belongs to a particular class. Naive Bayes classifier is the simplest Bayesian classifier, which is based on Bayes theory with strong (naive) assumption that all variables contribute toward classification and are mutually correlated. A Bayesian network is another classifier, it is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph, where the nodes are in one-to one correspondence with the features.

Another examples of the graphical models are Hidden Markov Model (HMM) Rabiner (1989) and Conditional Random Field (CRF) Lafferty et al. (2001). The former is a generative model and it gives the output directly by modeling the transition matrix based on the training data. It assumes that the system being modeled is a Markov process with unobserved (hidden) states. The latter is a discriminative model which outputs a confidence measure. It can be considered as a generalization of HMM.

Instance based learners is another category of classifiers. One of the simplest classifier is the k-Nearest Neighbour (k-NN) Cover and Hart (1967). It locates the k nearest instances to the given query instance and determines its label by selecting the single most frequent label of nearest instances. The main limitation of this classifier is that it requires to store all the instances and it is sensitive to the choice of the similarity function to compare instances.

Moreover, there is general agreement that it is very sensitive to irrelevant features. There are many variants of the k-NN algorithm, e.g. CNN and UNN. Condensed nearest neighbor (CNN) Hart (1968) is designed to reduce the data set for classification. It selects the set of prototypes from the training data to classify samples almost as accurately as the nearest neighbour with the whole data set. Another variant of the k-NN algorithm is the Universal Nearest Neighbors Piro et al. (2010), which is a boosting algorithm for inducing a leveraged k-NN rule. This rule generalizes the k-NN to weighted voting, i.e. the votes of nearest neighbors are weighted by means of real coefficients, where the weights (called leveraging coefficients) are iteratively learned from training data.

Decision trees Murthy (1998) belong to logic category of classifiers. Decision trees are trees, which classify instances by sorting them based on feature values. Each node in a decision tree represents a test on a feature, each branch represents an outcome of the test, and each leaf node represents the class label. There are several measures for finding the best features for the construction of a decision tree: Information gain, Gain ratio, Gini index, ReliefF algorithm, Chi square, and others. However, no measure is significantly better than others. The construction of the optimal decision tree is an NP-complete problem. The popular decision tree algorithms are: Iterative Dichotomiser 3 (ID3), C4.5, and Classification and Regression Trees (CART). The main limitation of the decision trees is that they tend to overfit the training data. Random forest classifier (Breiman (2001), Genuer et al. (2008)) solves this problem. It uses a multitude of decision trees and outputs the class label based on the votes from all the individual decision trees. Moreover, it uses a random selection of features to split each node.

Another category of classifiers is perceptron based techniques. Artificial Neural Networks (ANNs) (Rumelhart et al. (1985), Zhang (2000)) are multi-layer neural networks, which consist of a number of connected units (neurons). ANNs consist of three types of layers: input layer with input units which receive information to be processed, output layer with output units which give the result of the algorithm, and hidden layers with hidden units which process the data. An ANN learns the weights of the connections between neurons in order to determine the mapping between the input and the output. They are many types of ANNs: single layer perceptron, RBF network, DNNs, CNNs, and others. A single layer perceptron is the simpliest neural network based on a linear combination of a set of weights with the feature vector. A Deep Neural Network (DNN) is a neural network with at least one hidden layer of units between the input and output layers. A Radial Basis Function (RBF) is a three-layer feedback network, in which each hidden unit implements a radial activation function and each output unit applies a weighted sum of hidden units outputs.

A Convolutional Neural Network (CNN) is another type of a neural network that can be directly applied on the raw input, thus automating the process of feature construction. Boosting Schapire (1999) is a machine learning meta algorithm, which creates a strong classifier from a set of weak classifiers. The algorithm iteratively learns weak classifiers and adds them to a final strong classifier with weights which are typically corresponding to the accuracy. After a weak classifier is added, the data is reweighted, and typically the correctly classified samples loose weight, and the misclassified samples gain weight so the boosting algorithm will focus on them in the next iteration step. A weak classifier is defined as a classifier which works at least as well as a random classifier. A strong classifier should be well correlated with the true classification. The popular boosting algorithms are: AdaBoost, GentleBoost, BrownBoost, LogitBoost, Bootstraping, and others.

Support Vector Machines (SVMs) (Vapnik and Vapnik (1998), Burges (1998), Cristianini and Shawe-Taylor (2000)) belong to another category of classifiers. They are maximizing the distance between a hyperplane that separates two classes of data and instances on either side of it. They can perform linear separation and also non-linear separation using a kernel function. Moreover, they reach the global minimum and avoid ending in a local minimum, what may happen in other search algorithms such as neural networks. Finally, they typically provide very good results.

Many of the above classification techniques have been successfully applied to action recognition in videos, e.g. HMMs Yamato et al. (1992), k-NN (Efros et al. (2003), Blank et al. (2005), Thurau and Hlaváč (2008)), ANNs Iosifidis et al. (2012), CNNs Karpathy et al. (2014), Boosting (Nowozin et al. (2007), Fathi and Mori (2008)), and SVMs (Dollár et al. (2005), Laptev (2005), Laptev et al. (2008), Liu and Shah (2008), Wang et al. (2011a)). Over the last years, SVMs is the most popular classification technique used in action recognition in videos. All the above classification algorithms have pros and cons and we refer to the work of Kotsiantis Kotsiantis et al. (2007) for the details. According to that work, SVMs achieve the best accuracy in general, in comparison with the Decision Trees, Neural Networks, Nave Bayes, k-NN, and Rule-learners. They are also at least as good as others in speed of classification, tolerance to irrelevant attributes, tolerance to redundant attributes, and tolerance to highly interdependent attributes. However, there is no single learning algorithm that can uniformly outperform other techniques over all datasets. SVMs have a sound theoretical foundation, and they are considered as a must try Wu et al. (2008) as they are one of the most robust and accurate methods. However, SVMs also have cons Kotsiantis et al. (2007), e.g. their performance highly relies on the selection of an appropriate kernel function, they have low speed of learning w.r.t. the number of attributes and the number of instances, and they do not handle well model parameters. For action recognition in videos, SVMs are the most widely used supervised learning classifiers. They achieve very good results, there exist kernel functions that give good results, the number of instances and the number of attributes are typically not large (up to several thousands), and there are typically not many classifier parameters to learn.

Ensemble of Classifiers The above supervised learning techniques use an individual method to perform a classification. Another type of approaches creates an ensemble of classifiers to obtain better predictive performance. Over the last years, numerous methods have been proposed for that (Kotsiantis et al. (2007), Dietterich (2000), Rokach (2010)), and these methods typically use: (a) various subsets of training data with a single learning approach, (b) various training parameters with a single training approach, and/or (c) various learning approaches. Although many ensemble methods have been proposed, there is no clear picture which technique is the best (Kotsiantis et al. (2007), Vilalta and Drissi (2002)). An ensemble of classifiers have been used by several action recognition approaches, e.g. (Yang and Shah (2012), Izadinia and Shah (2012), Oh et al. (2014)). Finding the right ensemble method is still an open machine learning research problem. An ensemble of classifiers may improve results for some features, techniques, and decrease results for others.

4. Techniques Using Weakly-Supervised Classifiers

instead of supervised action recognition, it uses different unsupervised approaches to define long-term activity models. It also benifits from some supervised information as hints to help reach better models.

??? Tinne can elaborate on this section ???

- Using Visual Information and Text on Large Dataset
- Using Visual Information and Audio

- Using Visual Information, Text, Audio and etc.

5. Techniques Using Unsupervised Classifiers

From the very beginning, supervised approaches has been one of the most popular approaches for recognizing human actions Aggarwal and Cai (1999). Recently, a particular attention has been drawn on extracting action descriptors using space-time interest points, local image descriptors and bag-of-words (BoW) approach Laptev et al. (2008); Wang et al. (2011a). For simple and short-term actions such as walking, hand waving, these approaches report high recognition rates. For long-term activities, there are many unsupervised approaches that model the global motion pattern and detect abnormal events by finding the trajectories that do not fit in the pattern. Many methods has been applied on traffic surveillance videos to learn the regular traffic dynamics (e.g., cars passing a cross road) and detect abnormal patterns (e.g., a pedestrian crossing the road) Hu et al. (2006).

Activities of daily living (ADL), such as cooking, consists of long-term complex activities that are composed of short-term actions. Supervised approaches only provides a representation of short-term local body movements. Therefore, they require too much user interaction: splitting long videos into clips that contains only one simple action and labeling this very large amount of clipped data. Since the beginning and the end of the activity is not known, it is hard to train classifiers that can distinguish long-term actions. On the other hand, with existing unsupervised approaches, modeling the global motion pattern cannot capture the complex structure of long-term human activities.

many supervised approaches have been proposed for recognizing human actions from videos. Different features has been examined for robust and discriminative representation of actions(section 4). In addition, many machine learning approaches has been applied to model the actions and obtain robust classifiers. Lots of methods rely on skeleton detection. However, skeleton detection is noisy when the view is not frontal or there is occlusion. In all of these methods, the common approach is to use datasets that include short, simple and well-clipped actions. Features of interest are extracted from the huge set of short clips and labeled. Then, using this huge amount of labeled data, a supervised classifier is trained to learn the model of each action. Benefiting from the very well-organized training sets, many approaches achieve state-of-the-art results. However, in the case of ADL, such approaches cannot handle the complexity and provide a discriminative representation of actions.

By addressing this disadvantage of supervised methods, there are unsupervised methods that directly learn activity models from the whole data (videos). Hu et al. Hu et al. (2006) learn motion patterns in traffic surveillance videos by using a two-layered trajectory clustering via fuzzy k-means algorithm: clustering first in space and second in time. This idea has been extended in Morris and Trivedi (2011) and a three-layered clustering is performed on trajectories in order to learn the variations: first in spatial routes, second in time duration, and third in speed. Then, the spatio-temporal dynamics of each cluster is encoded by training HMMs using the most representative examples of clusters. In Calderara et al. (2007), normal motion is modeled as a mixture of Von Mises distributions and parameters of mixture distributions are learned by clustering direction information of trajectories. The trajectories that do not fit the model are classified as abnormal. Bobick and Wilson (1997) use dynamic programming based approaches to classify activities. These methods are only effective when constraints on time ordering hold. Similarly, the approach in Dee et al. (2012) builds semantic scene models by clustering trajectory points and motion direction. They segment the regions in the scene that are similar in terms of space and motion direction. The approach in Porikli (2004) uses HMM to represent trajectory paths by clustering and captures spatio-temporal patterns in trajectory paths. Clustering is based on finding the number of clusters by checking how well eigenvectors of the trajectory correlation matrix span the subspace. These approaches allow high-level analysis of activities for detecting abnormalities in traffic videos. However, ADLs are more complex than traffic flow. Therefore, using only global object trajectories will be insufficient to capture spatio-temporal modalities of ADL and discriminate among them (e.g., there will be no difference between “standing next to table” and “eating at the table”).

Another trajectory-based approach for human activity recognition Gao and Sun (2013) uses HDP to address the limitation of HMMs regarding the prediction of number of human motion states. For each activity, the distribution of hidden motion labels, mean and covariance parameters, and transition probabilities are sampled iteratively via Gibbs sampling. Thus, it requires too much computation to obtain the number of motion states. Furthermore, there are many approaches in the field of computer vision based assistive technologies that focus on assisting elderly people while performing an ADL (e.g. washing hand) Hoey et al. (2012). In Hoey et al. (2010), Hoey et al. introduces a system that assists people with mild to moderate dementia in the task of hand washing. The system tracks the two hands and objects (e.g. towel) and utilizes a Partially Observable Markov Decision Process (POMDP) for planning and decision making. The POMDP includes a set of eight steps of hand washing, a set of six simple prompts and models the user's responsiveness, awareness, and overall dementia level.

Peters et al. in Peters et al. (2014) proposes an assistive system that support people with moderate cognitive disabilities in the execution of brushing teeth. A color-based object detector is used to find the locations of objects involved (e.g. brush, paste), a Bayesian network classifier is used to handle the variabilities in behavior recognition, and maintain an ordering constraint graph to model a set of ordering constraints between user behaviors.

6. Generating Complex Description: Verbalize, Machine Translation

??? Claire can elaborate on this section ???

- TACOS Cooking Dataset

References

Aggarwal, J., Cai, Q., 1999. Human motion analysis: A review. Computer Vision and Image Understanding 73, 428 – 440. URL: http://www.sciencedirect.com/science/article/pii/S1077314298907445, doi:http://dx.doi.org/10.1006/cviu.1998.0744.

Ali, S., Basharat, A., Shah, M., 2007. Chaotic invariants for human action recognition, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE. pp. 1–8.

Avila, S., Thome, N., Cord, M., Valle, E., Araujo, A.d.A., 2013. Pooling in image representation: The visual codeword point of view. Computer Vision and Image Understanding 117, 453–465.

Avila, S., Thome, N., Cord, M., Valle, E., et al., 2011. Bossa: Extended bow formalism for image classification, in: Image Processing (ICIP), 2011 18th IEEE International Conference on, IEEE. pp. 2909–2912.

Banerjee, P., Nevatia, R., 2011. Learning neighborhood cooccurrence statistics of sparse features for human activity recognition, in: Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on, IEEE. pp. 212–217.

Bay, H., Tuytelaars, T., Van Gool, L., 2006. Surf: Speeded up robust features, in: Computer vision–ECCV 2006. Springer, pp. 404–417.

Beaudet, P.R., 1978. Rotationally invariant image operators, in: International Joint Conference on Pattern Recognition, p. 583.

Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R., 2005. Actions as space-time shapes, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, IEEE. pp. 1395–1402.

Bobick, A.F., Davis, J.W., 2001. The recognition of human movement using temporal templates. Pattern Analysis and Machine Intelligence, IEEE Transactions on 23, 257–267.

Bobick, A.F., Wilson, A.D., 1997. A state-based approach to the representation and recognition of gesture. Pattern Analysis and Machine Intelligence, IEEE Transactions on 19, 1325–1337.

Breiman, L., 2001. Random forests. Machine learning 45, 5–32.

Brendel, W., Fern, A., Todorovic, S., 2011. Probabilistic event logic for interval-based event recognition, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3329–3336. doi:10.1109/CVPR.2011.5995491.

Burges, C.J., 1998. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery 2, 121–167.

Calderara, S., Cucchiara, R., Prati, A., 2007. Detection of abnormal behaviors using a mixture of von mises distributions, in: IEEE Conference on Advanced Video and Signal Based Surveillance, 2007. AVSS 2007., IEEE. pp. 141–146.

Cao, Y., Tao, L., Xu, G., 2009. An event-driven context model in elderly health monitoring. Ubiquitous, Autonomic and Trusted Computing, Symposia and Workshops on , 120–124doi:http://doi.ieeecomputersociety.org/10.1109/UIC-ATC.2009.47.

Chatfield, K., Lempitsky, V.S., Vedaldi, A., Zisserman, A., 2011. The devil is in the details: an evaluation of recent feature encoding methods., in: BMVC, p. 8.

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. Information Theory, IEEE Transactions on 13, 21–27.

Cristianini, N., Shawe-Taylor, J., 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge university press.

Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints, in: Workshop on statistical learning in computer vision, ECCV, Prague. pp. 1–2.

Cula, O.G., Dana, K.J., 2001. Compact representation of bidirectional texture functions, in: Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, IEEE. pp. I–1041.

Dalal, N., Triggs, B., 2005. Histograms of oriented gradients for human detection, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, IEEE. pp. 886–893.

Dalal, N., Triggs, B., Schmid, C., 2006. Human detection using oriented histograms of flow and appearance, in: Computer Vision–ECCV 2006. Springer, pp. 428–441.

Dee, H.M., Cohn, A.G., Hogg, D.C., 2012. Building semantic scene models from unconstrained video. Computer Vision and Image Understanding 116, 446 – 456. URL: http://www.sciencedirect.com/science/

article/pii/S1077314211002025, doi:http://dx.doi.org/10.1016/j.cviu.2011.09.005. special issue on Semantic Understanding of Human Behaviors in Image Sequences.

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine learning 40, 139–157.

Dollár, P., Rabaud, V., Cottrell, G., Belongie, S., 2005. Behavior recognition via sparse spatio-temporal features, in: Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on, IEEE. pp. 65–72.

Efros, A., Berg, A.C., Mori, G., Malik, J., et al., 2003. Recognizing action at a distance, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE. pp. 726–733.

Fathi, A., Mori, G., 2008. Action recognition by learning mid-level motion features, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Gao, Q., Sun, S., 2013. Trajectory-based human activity recognition with hierarchical dirichlet process hidden markov models, in: Signal and Information Processing (ChinaSIP), 2013 IEEE China Summit & International Conference on, IEEE. pp. 456–460.

van Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Smeulders, A.W., 2008. Kernel codebooks for scene categorization, in: Computer Vision–ECCV 2008. Springer, pp. 696–709.

Genuer, R., Poggi, J.M., Tuleau, C., 2008. Random forests: some methodological insights. arXiv preprint arXiv:0811.3619 .

Gilbert, A., Illingworth, J., Bowden, R., 2009. Fast realistic multi-action recognition using mined dense spatio-temporal features, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE. pp. 925–931.

Harris, C., Stephens, M., 1988. A combined corner and edge detector., in: Alvey vision conference, Citeseer. p. 50.

Hart, P., 1968. The condensed nearest neighbor rule (corresp.). Information Theory, IEEE Transactions on 14, 515–516. doi:10.1109/TIT.1968.1054155.

Hoey, J., Boutilier, C., Poupart, P., Olivier, P., Monk, A., Mihailidis, A., 2012. People, sensors, decisions: Customizable and adaptive technologies for assistance in healthcare. ACM Transactions on Interactive Intelligent Systems (TiiS) 2, 20.

Hoey, J., Poupart, P., von Bertoldi, A., Craig, T., Boutilier, C., Mihailidis, A., 2010. Automated handwashing assistance for persons with dementia using video and a partially observable markov decision process. Computer Vision and Image Understanding 114, 503–519.

Hu, W., Xiao, X., Fu, Z., Xie, D., Tan, T., Maybank, S., 2006. A system for learning statistical motion patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1450–1464.

Iosifidis, A., Tefas, A., Pitas, I., 2012. View-invariant action recognition based on artificial neural networks. Neural Networks and Learning Systems, IEEE Transactions on 23, 412–424.

Izadinia, H., Shah, M., 2012. Recognizing complex events using large margin joint low-level event model, in: Proceedings of the 12th European Conference on Computer Vision - Volume Part IV, Springer-Verlag, Berlin, Heidelberg. pp. 430–444. URL: http://dx.doi.org/10.1007/978-3-642-33765-9_31, doi:10.1007/978-3-642-33765-9_31.

Jain, M., Jégou, H., Bouthemy, P., 2013. Better exploiting motion for better action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE. pp. 2555–2562.

Jansson, G., Johansson, G., 1973. Visual perception of bending motion. Perception 2, 321–326.

Jégou, H., Douze, M., Schmid, C., Pérez, P., 2010. Aggregating local descriptors into a compact image representation, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 3304–3311.

Jégou, H., Perronnin, F., Douze, M., Sanchez, J., Perez, P., Schmid, C., 2012. Aggregating local image descriptors into compact codes. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34, 1704–1716.

Jensen, F.V., 1996. An introduction to Bayesian networks. volume 210. UCL press London.

Kaaniche, M.B., Brémond, F., 2009. Tracking hog descriptors for gesture recognition, in: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on, IEEE. pp. 140–145.

Kadir, T., Brady, M., 2003. Scale saliency: A novel approach to salient feature and scale selection .

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, IEEE. pp. 1725–1732.

Khoshelham, K., Elberink, S.O., 2012. Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 12, 1437–1454.

Klaser, A., Marszałek, M., Schmid, C., 2008. A spatio-temporal descriptor based on 3d-gradients, in: BMVC 2008-19th British Machine Vision Conference, British Machine Vision Association. pp. 275–1.

Kotsiantis, S.B., Zaharakis, I., Pintelas, P., 2007. Supervised machine learning: A review of classification techniques.

Kovashka, A., Grauman, K., 2010. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 2046–2053.

Krapac, J., Verbeek, J., Jurie, F., 2011. Modeling spatial layout with fisher vectors for image categorization, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE. pp. 1487–1494.

Kumar, P., Ranganath, S., Weimin, H., Sengupta, K., 2005. Framework for real-time behavior interpretation from traffic video. Intelligent Transportation Systems, IEEE Transactions on 6, 43–53. doi:10.1109/TITS.2004.838219.

Kwak, S., Han, B., Han, J.H., 2011. Scenario-based video event recognition by constraint flow., in: CVPR, IEEE. pp. 3345–3352. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2011.html#KwakHH11.

Lafferty, J., McCallum, A., Pereira, F.C., 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

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

Laptev, I., Lindeberg, T., 2006. Local descriptors for spatio-temporal recognition, in: Spatial Coherence for Visual Motion Analysis. Springer, pp. 91–103.

Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B., 2008. Learning realistic human actions from movies, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Lavee, G., Rivlin, E., Rudzsky, M., 2009. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 39, 489–504. doi:10.1109/TSMCC.2009.2023380.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, IEEE. pp. 2169–2178.

Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y., 2011. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, Washington, DC, USA. pp. 3361–3368. URL: http://dx.doi.org/10.1109/CVPR.2011.5995496, doi:10.1109/CVPR.2011.5995496.

Litomisky, K., 2012. Consumer rgb-d cameras and their applications. Rapport technique, University of California , 20.

Liu, J., Shah, M., 2008. Learning human actions via information maximization, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60, 91–110.

Lucas, B.D., Kanade, T., et al., 1981. An iterative image registration technique with an application to stereo vision., in: IJCAI, pp. 674–679.

Lv, F., Song, X., Wu, B., Kumar, V., Nevatia, S.R., 2006. Left luggage detection using bayesian inference, in: In PETS.

Matikainen, P., Hebert, M., Sukthankar, R., 2009. Trajectons: Action recognition through the motion analysis of tracked features, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, IEEE. pp. 514–521.

Matikainen, P., Hebert, M., Sukthankar, R., 2010. Representing pairwise spatial and temporal relations for action recognition, in: Computer Vision–ECCV 2010. Springer, pp. 508–521.

Messing, R., Pal, C., Kautz, H., 2009. Activity recognition using the velocity histories of tracked keypoints, in: Computer Vision, 2009 IEEE 12th International Conference on, IEEE. pp. 104–111.

Morris, B., Trivedi, M., 2011. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 33, 2287–2301. doi:10.1109/TPAMI.2011.64.

Murthy, S.K., 1998. Automatic construction of decision trees from data: A multi-disciplinary survey. Data mining and knowledge discovery 2, 345–389.

Niebles, J.C., Wang, H., Fei-Fei, L., 2008. Unsupervised learning of human action categories using spatial-temporal words. International journal of computer vision 79, 299–318.

Nowozin, S., Bakir, G., Tsuda, K., 2007. Discriminative subsequence mining for action classification, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE. pp. 1–8.

Oh, S., McCloskey, S., Kim, I., Vahdat, A., Cannons, K.J., Hajimirsadeghi, H., Mori, G., Perera, A.A., Pandey, M., Corso, J.J., 2014. Multimedia event detection with multimodal feature fusion and temporal concept localization. Machine vision and applications 25, 49–69.

Oikonomopoulos, A., Patras, I., Pantic, M., 2005. Spatiotemporal salient points for visual recognition of human actions. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on 36, 710–719.

Oneata, D., Verbeek, J., Schmid, C., 2013. Action and event recognition with fisher vectors on a compact feature set, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE. pp. 1817–1824.

Oshin, O., Gilbert, A., Bowden, R., 2011. Capturing the relative distribution of features for action recognition, in: Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, IEEE. pp. 111–116.

Park, S., Aggarwal, J.K., 2004. A hierarchical bayesian network for event recognition of human actions and interactions. Multimedia Syst. 10, 164–179. URL: `http://dblp.uni-trier.de/db/journals/mms/mms10.html#ParkA04`.

Perronnin, F., Sánchez, J., Mensink, T., 2010. Improving the fisher kernel for large-scale image classification, in: Computer Vision–ECCV 2010. Springer, pp. 143–156.

Peters, C., Hermann, T., Wachsmuth, S., Hoey, J., 2014. Automatic task assistance for people with cognitive disabilities in brushing teeth-a user study with the tebra system. ACM Transactions on Accessible Computing (TACCESS) 5, 10.

Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A., 2008. Lost in quantization: Improving particular object retrieval in large scale image databases, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Piro, P., Nock, R., Nielsen, F., Barlaud, M., 2010. Boosting k-nn for categorization of natural scenes. arXiv preprint arXiv:1001.1221 .

Porikli, F., 2004. Learning object trajectory patterns by spectral clustering, in: IEEE International Conference on Multimedia and Expo, 2004. ICME'04. 2004, IEEE. pp. 1171–1174.

Rabiner, L.R., 1989. A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77, 257–286.

Raptis, M., Kirovski, D., Hoppe, H., 2011. Real-time classification of dance gestures from skeleton animation, in: Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation, ACM. pp. 147–156.

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intelligence Review 33, 1–39.

Romdhane, R., Crispim, C., Bremond, F., Thonnat, M., 2013. Activity recognition and uncertain knowledge in video scenes, in: Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on, pp. 377–382. doi:10.1109/AVSS.2013.6636669.

Rosten, E., Drummond, T., 2006. Machine learning for high-speed corner detection, in: Computer Vision–ECCV 2006. Springer, pp. 430–443.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1985. Learning internal representations by error propagation. Technical Report. DTIC Document.

Ryoo, M.S., Aggarwal, J.K., 2006. Recognition of composite human activities through context-free grammar based representation., in: CVPR (2), IEEE Computer Society. pp. 1709–1718. URL: http://dblp.uni-trier.de/db/conf/cvpr/cvpr2006-2.html#RyooA06.

Ryoo, M.S., Aggarwal, J.K., 2009a. Semantic representation and recognition of continued and recursive human activities. International Journal of Computer Vision 82, 1–24. URL: http://dblp.uni-trier.de/db/journals/ijcv/ijcv82.html#RyooA09.

Ryoo, M.S., Aggarwal, J.K., 2009b. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in: Computer vision, 2009 ieee 12th international conference on, IEEE. pp. 1593–1600.

Salton, G., 1968. Automatic information organization and retrieval .

Schapire, R.E., 1999. A brief introduction to boosting, in: Ijcai, pp. 1401–1406.

Schüldt, C., Laptev, I., Caputo, B., 2004. Recognizing human actions: a local svm approach, in: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on, IEEE. pp. 32–36.

Scovanner, P., Ali, S., Shah, M., 2007. A 3-dimensional sift descriptor and its application to action recognition, in: Proceedings of the 15th international conference on Multimedia, ACM. pp. 357–360.

Shi, J., Tomasi, C., 1994. Good features to track, in: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on, IEEE. pp. 593–600.

Shotton, J., Girshick, R., Fitzgibbon, A., Sharp, T., Cook, M., Finocchio, M., Moore, R., Kohli, P., Criminisi, A., Kipman, A., et al., 2013. Efficient human pose estimation from single depth images. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35, 2821–2840.

Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T., 2005. Discovering object categories in image collections .

Sivic, J., Zisserman, A., 2003. Video google: A text retrieval approach to object matching in videos, in: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE. pp. 1470–1477.

Sun, C., Nevatia, R., 2013. Large-scale web video event classification by use of fisher vectors, in: Applications of Computer Vision (WACV), 2013 IEEE Workshop on, IEEE. pp. 15–22.

Sun, J., Mu, Y., Yan, S., Cheong, L.F., 2010. Activity recognition using dense long-duration trajectories, in: Multimedia and Expo (ICME), 2010 IEEE International Conference on, IEEE. pp. 322–327.

Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.S., Li, J., 2009. Hierarchical spatio-temporal context modeling for action recognition, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 2004–2011.

Ta, A.P., Wolf, C., Lavoue, G., Baskurt, A., Jolion, J.M., 2010. Pairwise features for human action recognition, in: Pattern Recognition (ICPR), 2010 20th International Conference on, IEEE. pp. 3224–3227.

Thurau, C., Hlaváč, V., 2008. Pose primitive based human action recognition in videos or still images, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE. pp. 1–8.

Tomasi, C., Kanade, T., 1991. Detection and tracking of point features. School of Computer Science, Carnegie Mellon Univ. Pittsburgh.

Tran, S.D., Davis, L.S., 2008. Event modeling and recognition using markov logic networks, in: ECCV '08: Proceedings of the 10th European Conference on Computer Vision, Springer-Verlag, Berlin, Heidelberg. pp. 610–623. doi:http://dx.doi.org/10.1007/978-3-540-88688-4_45.

Vapnik, V.N., Vapnik, V., 1998. Statistical learning theory. volume 1. Wiley New York.

Vilalta, R., Drissi, Y., 2002. A perspective view and survey of meta-learning. Artificial Intelligence Review 18, 77–95.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2011a. Action recognition by dense trajectories, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 3169–3176.

Wang, H., Klaser, A., Schmid, C., Liu, C.L., 2011b. Action recognition by dense trajectories, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, pp. 3169–3176. doi:10.1109/CVPR.2011.5995407.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2013. Dense trajectories and motion boundary descriptors for action recognition. International journal of computer vision 103, 60–79.

Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C., 2009. Evaluation of local spatio-temporal features for action recognition, in: BMVC 2009-British Machine Vision Conference, BMVA Press. pp. 124–1.

Wang, J., Chen, Z., Wu, Y., 2011c. Action recognition with multiscale spatio-temporal contexts, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE. pp. 3185–3192.

Wang, J., Liu, Z., Wu, Y., Yuan, J., 2012. Mining actionlet ensemble for action recognition with depth cameras, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE. pp. 1290–1297.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y., 2010. Locality-constrained linear coding for image classification, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE. pp. 3360–3367.

Willems, G., Tuytelaars, T., Van Gool, L., 2008. An efficient dense and scale-invariant spatio-temporal interest point detector, in: Computer Vision–ECCV 2008. Springer, pp. 650–663.

Wong, S.F., Cipolla, R., 2007. Extracting spatiotemporal interest points using global information, in: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE. pp. 1–8.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Philip, S.Y., et al., 2008. Top 10 algorithms in data mining. Knowledge and Information Systems 14, 1–37.

Yamato, J., Ohya, J., Ishii, K., 1992. Recognizing human action in time-sequential images using hidden markov model, in: Computer Vision and Pattern Recognition, 1992. Proceedings CVPR'92., 1992 IEEE Computer Society Conference on, IEEE. pp. 379–385.

Yang, Y., Shah, M., 2012. Complex events detection using data-driven concepts, in: Computer Vision–ECCV 2012. Springer, pp. 722–735.

Yilma, A., Shah, M., 2005. Recognizing human actions in videos acquired by uncalibrated moving cameras, in: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, IEEE. pp. 150–157.

Zaidenberg, S., Boulay, B., Brmond, F., 2012. A generic framework for video understanding applied to group behavior recognition. CoRR abs/1206.5065. URL: `http://dblp.uni-trier.de/db/journals/corr/corr1206.html#abs-1206-5065`.

Zhang, G.P., 2000. Neural networks for classification: a survey. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 30, 451–462.

Zhou, Q., Wang, G., Jia, K., Zhao, Q., 2013. Learning to share latent tasks for action recognition, in: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE. pp. 2264–2271.

Zhou, X., Yu, K., Zhang, T., Huang, T.S., 2010. Image classification using super-vector coding of local image descriptors, in: Computer Vision–ECCV 2010. Springer, pp. 141–154.

Zouba, N., Bremond, F., Thonnat, M., 2010. An activity monitoring system for real elderly at home: Validation study, in: Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on, pp. 278–285. doi:10.1109/AVSS.2010.83.