A survey of vision-based methods for action representation, segmentation and recognition

Daniel Weinland, Rémi Ronfard, Edmond Boyer

A Survey of Vision-Based Methods for Action Representation, Segmentation and Recognition

Daniel Weinland[a], Remi Ronfard[b], Edmond Boyer[c]

[a] *Deutsche Telekom Laboratories, TU-Berlin*
Berlin, Germany
[b] *INRIA - Team Lear*
Grenoble, France
[c] *INRIA - Team Perception*
Grenoble, France

Abstract

Action recognition has become a very important topic in computer vision, with many fundamental applications, in robotics, video surveillance, human computer interaction, and multimedia retrieval among others and a large variety of approaches have been described. The purpose of this survey is to give an overview and categorization of the approaches used. We concentrate on approaches that aim on classification of full-body motions, such as kicking, punching, waving, etc. and we categorize them according to how they represent the spatial and temporal structure of actions; how they segment actions from an input stream of visual data; and how they learn a view-invariant representation of actions.

Key words: action/activity recognition, survey, computer vision

1. Introduction

Action recognition is a very active research topic in computer vision with many important applications, including human-computer interfaces, content-based video indexing, video surveillance, and robotics, among others. Historically, visual action recognition has been divided into sub-topics such as gesture recognition for human-computer interfaces [27, 88], facial expression recognition [152], and movement behavior recognition for video surveillance [44]. However full-body actions usually include different motions and

require a unified approach for recognition, encompassing facial actions, hand actions and feet actions.

Action recognition is the process of naming actions, usually in the simple form of an action verb, using sensory observations. Technically, an action is a sequence of movements generated by a human agent during the performance of a task. As such, it is a four-dimensional object, which may be further decomposed into spatial and temporal *parts*. In this paper, we are only concerned with visual observations, typically by means of one or more video cameras, but it should be noted that actions can of course also be recognized from other sensory channels, including audio. An action label is a name, such that an average human agent can understand and perform the named action. The task of action recognition is to name actions, i.e. determine the action label that best describes an action instance, even when performed by different agents under different viewpoints, and in spite of large differences in manner and speed. A typical set-up for testing and evaluating action recognition systems consist in sending instructions to the actors, using simple action verb imperatives, and to compare them with the recognized action names. Figure 1 illustrates the major components of a generic action recognition system and their typical arrangement.

Feature extraction is the main vision task in action recognition and consist in extracting posture and motion cues from the video that are discriminative with respect to human actions. Very different representations can be used, ranging from complex body models to simple silhouette images. In either case, issues such as person location, robustness to partial occlusion, background clutter, shadows and different illumination need to be addressed. Further representations should provide some insensitivity to different types of clothing and physiques.

Action learning and classification are the steps of learning statistical models from the extracted features, and using those models to classify new feature observations. A major challenge thereby is to deal with the large variability that an action class can exhibit, in particular if performed by different subjects of different gender and size, and with different speed and style. Action categories which might seem clearly defined to us, such as kicking, punching, or waving, for instance, can have very large variability when performed in practice. It is thus a particular challenge to design an action model, which identifies for each action the characteristic attitudes, while maintaining appropriate adaptability to all forms of variations.

Action segmentation is necessary to cut streams of motions into single

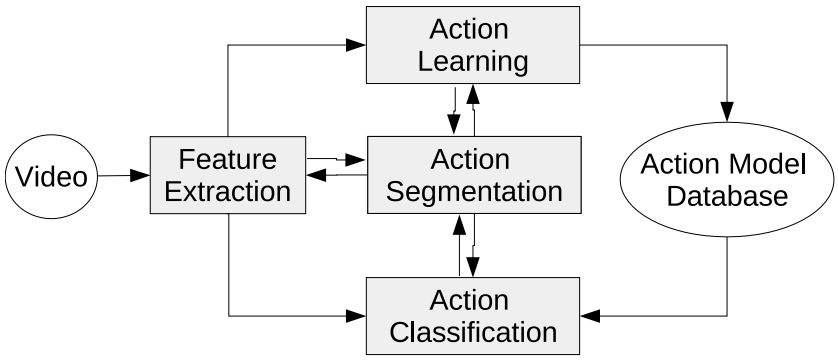


Figure 1: A typical data-flow for generic action recognition system comprises inter-dependent stages of feature extraction, learning, segmentation and classification.

action instances that are consistent to the set of initial training sequences used to learn the models.

Vision-based techniques for representing, segmenting and recognizing human actions can be classified according to many different criteria, e.g. the body parts involved (facial expressions, hand gestures, upper-body gestures, full-body gestures, etc.); the selected image features (interest points, landmarks, edges, optical flow, etc.); the class of statistical models used for learning and recognition (nearest neighbors, discriminant analysis, Markov models, etc.). The classification we have found to be the most useful is how the different methods proposed in the literature represent the *spatial and temporal structure of actions*. Indeed, our analysis of the recent literature in computer vision reveals a large variety of approaches in both the temporal and the spatial dimensions, which can be summarized as follows. In the spatial domain, action recognition can be based on global image features, aligned to the geometry of the scene or camera; or on parametric image features, aligned to the geometry of the human body; or on statistical models describing the spatial distribution of local image features. We review those three important classes in Section 2. In the temporal domain, action recognition can be based on global temporal signatures, such as stacked features, that represent an entire action from start to finish; or on grammatical models that represent how the moments of actions are organized sequentially, usually with several states and transitions between those states; or as well on statistical models, describing distributions of possibly sparse an unstructured feature observations over time. We review those three important classes in Section

	Grammars	Templates	Temporal Statistics
Body Models	Body Grammars e.g. Ramanan[96], Green[36], Kitani[56], Lv[69], Wang[133], Guerra-Filho[39], Ikizler[45]	Body Templates e.g. Gavrila[33], Yacoob[143], Rao[97], Gritai[37]	Bag of Postures e.g. Ikizler[47]
Image Models	Image Grammar e.g. Elgammal[26], Ogale[85], Turaga[127], Weinland[136], Lv[70], Shi [116], Natarajan[80],	Image Template e.g. Bobick[7], Weinland[138], Laptev[61], Fathi[29], Souvenir[120], Farhadi[28]	Bag of Keyframes e.g. Carlsson[17], Efros[25], Weinland[135], Schindler[109]
Spatial Statistics	Space Bag of Trajectories, e.g. Messing[76]	Feature Template e.g. Laptev[59], Ke[55]	Bag of Events e.g. Schuldt[110], Boiman[9], Dollar[24], Niebles[81], Niebles[82], Klaser[57], Laptev[60]

Table 1: Classification of Action Recognition Methods based on Spatial (vertical axis) and Temporal Representations (horizontal axis). Only some of the more recent approaches are listed in each cell.

3. By combining the three main spatial classes with the three main temporal classes, we end up with a synoptic classification of action recognition into nine basic classes shown in Table 1.

The paper is organized as follows. First, we present a general overview of action recognition methods, based on how they represent the spatial structure of actions in Section 2, and the temporal structure of actions in Section 3. Then, we review the special topics of action segmentation in Section 4, view-invariance in Section 5 and experimental evaluations on publicly available datasets in Section 6.

2. Spatial Action Representations

We begin this survey with a review of spatial representation used to discriminate actions from visual data. As mentioned previously, a first step in

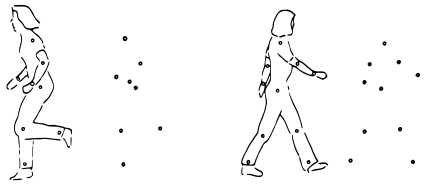


Figure 2: Illustration of moving light displays, taken from [50]. Johansson showed that humans can recognize actions merely from the motion of a few light displays attached to the human body. Awaiting publisher permission

action recognition is the extraction of image features that are discriminative with respect to posture and motion of the human body. Various representations have been suggested. They mainly contrast by the amount of high level information they represent versus how efficient they are to extract in practice. For the purpose of this survey, we classify them into three main groups: body models, image models, and local statistics.

2.1. Body models

In this section, we review methods that represent the spatial structure of actions with reference to the human body. In each frame of the observed video stream, the pose of a human body is recovered from a variety of available image features, and action recognition is performed based on such pose estimates. This is an intuitive and biologically-plausible approach to action recognition, which is supported by psychophysical work on visual interpretation of biological motion [50].

Johansson showed that humans can recognize actions merely form the motion of a few moving light displays (MLD) attached to the human body (Figure 2). Over several decades his experiments inspired approaches in action recognition, which used similar representations based on motion of landmark points on the human body. His experiments were also origin of the unresolved controversy on whether humans actually recognize actions directly from 2D motion patterns, or whether they first compute a 3D reconstruction from the motion of the patterns [122, 35]. In the context of machine vision, those two approaches have resulted in two main classes of methods [77]: 1) *recognition by reconstruction* of 3D body models and 2) *direct recognition* from 2D body models.

Recognition by reconstruction divides the task of action recognition in two well separate stages - a motion capture stage which estimate a 3D

model of the human body, typically represented as a kinematic joint model; and an action recognition stage which operates on joint trajectories. Two major difficulties are the large number of degrees-of-freedoms of the human body and the high variability of their shapes. As a result, a parametric model of the human body must be carefully selected and calibrated to support action recognition and generalization. A large variety of parametric models have been proposed over the years and we can only mention some of them. See Figure 3 for some examples.

In their early theoretical work on representation of three dimensional shapes [72], Marr and Nishihara proposed a body model consisting of a hierarchy of 3D cylindrical primitives, see Figure 3 a). Such a model was later adopted by several approaches, e.g. [42, 103]. More flexible body models based on super-quadrics have been used in [33], and models based on a textured spline model have been used in [36]. The approaches [96, 45] start from tracked patches in 2D and then lifts the 2D configurations into 3D, see Figure 3 c). Motion capture techniques requiring special markers attached to the human have also been used for action recognition, e.g. [16], see Figure 3 b). Other approaches directly work on the trajectories of 3D anatomical landmarks, e.g. head and hand trajectories [15, 12, 140].

Direct recognition approaches work from 2D models of the human body, i.e. labeled body parts, without lifting these into 3D. Common 2D representations are stick figures [40, 83], Figure 3 f), and 2D anatomical landmarks similar to Johansson's MLDs [35]. Other direct recognition approaches use coarse 2D body representations based on tracked blobs and patches, e.g. hand and head trajectories [121, 12], see Figure 3 d) or full body representations [143, 13].

To conclude this section, we should note that finding body parts and estimating parametric body models from images remains an unresolved problem, independent of the model used (2D or 3D). Even commercial MOCAP systems using special markers attached to the body rely on heavy user interaction, which makes them unsuitable for recognition tasks. Monocular, marker-less MOCAP, which is typically based on difficult non-convex optimizations, is highly prone to such issues as false initialization, convergence to local optima, or non-recovery from failure. Recent methods [104, 1] use strong prior learning to reduce such issues by assuming particular types of activities, walking or running for instance. Such prior models hence reduce the search space of possible poses considered, which however limits their application to action recognition [151, 90].

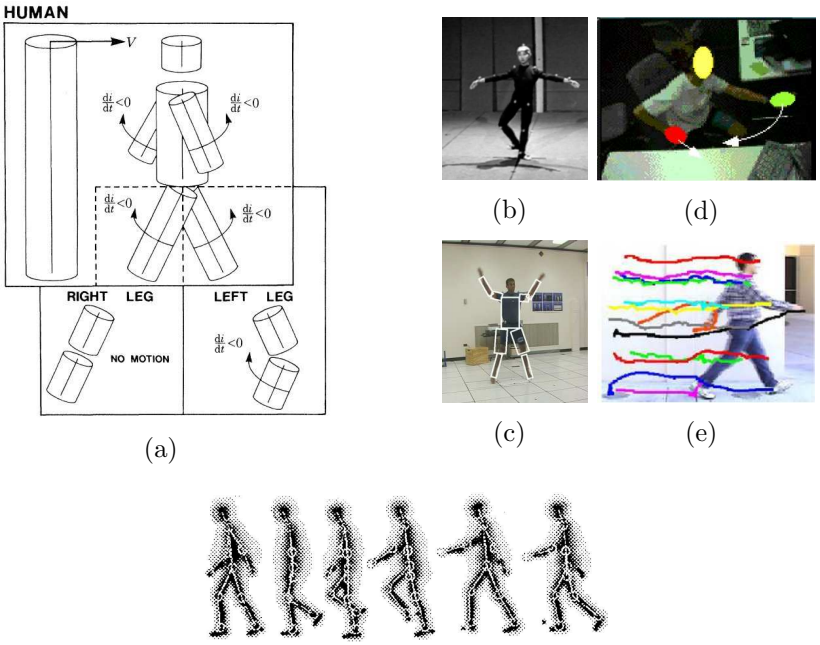


Figure 3: Body model representations: (a) hierarchical 3D model based on cylindrical primitives [73]; (b) ballet dancer with markers attached to body [16]; (c) body model based on rectangular patches [96]; (d) blob model [12]; (e) 2D marker trajectories [148]; (f) stick figure [40]. Awaiting publisher permission

2.2. Image models

In this section, we review global, image-based representations of actions, also sometimes called *holistic representations*, which do not require the detection and labeling of individual body parts. They only need to detect a region of interest (ROI) centered around the person. In most cases, features are then computed densely on a regular grid bounded by the detected region. As a general term, we call such a representation an *image model* of action. See Figure 4 for some examples.

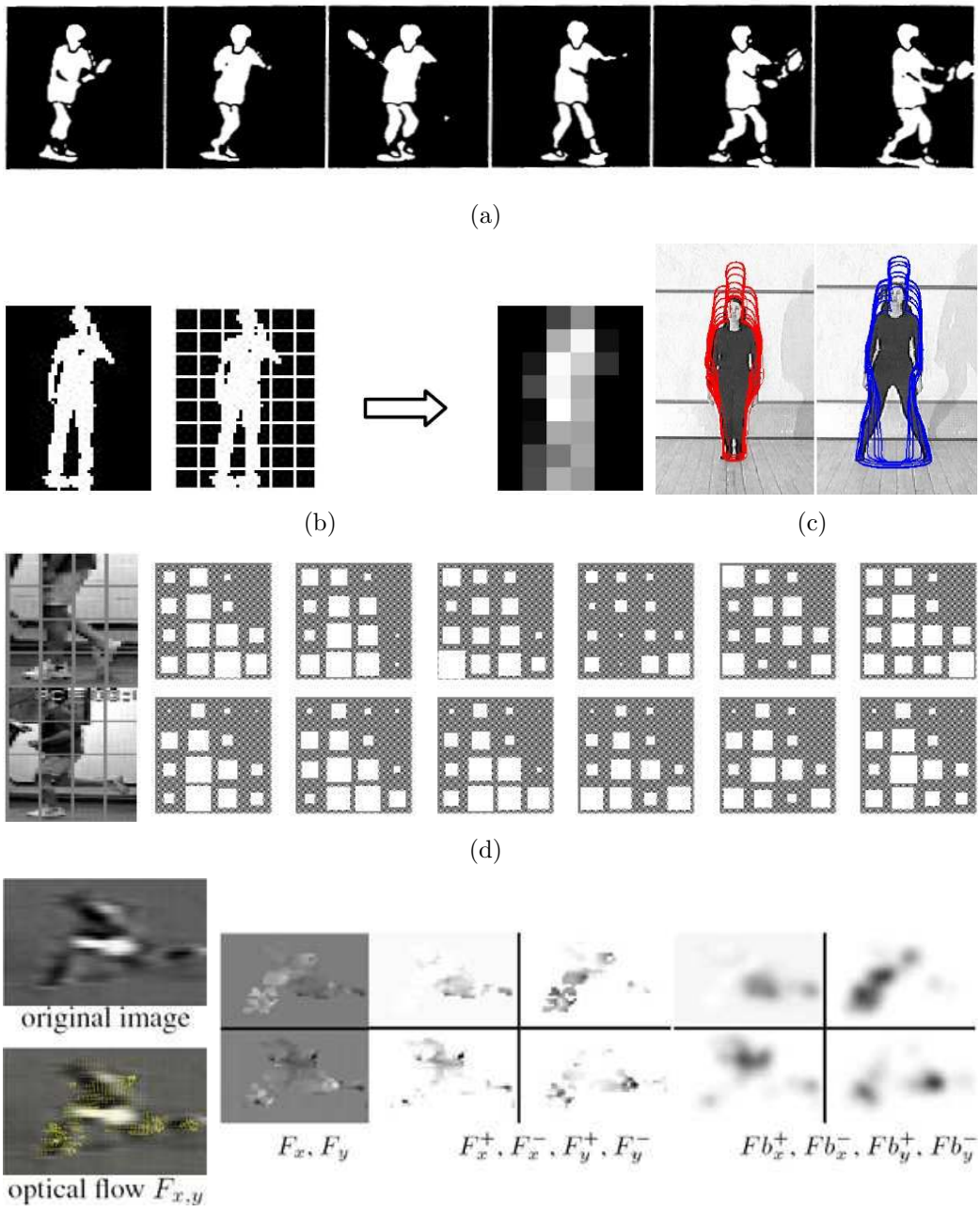
Image models can be much simpler than parametric body models. As a result, they can be computed more efficiently and robustly. Paradoxically, they have also been shown to be just as discriminative as body models with respect to many classes of actions.

A typical image model is presented by Darell et al. [23], where images of hand gestures are directly correlated, without feature extraction. Their work assumes however a static black background. In most other cases, background subtraction and feature extraction must be performed in a pre-processing stage.

An important class of image models uses silhouettes and contours of the human agent performing the action. As a good example, the seminal work on HMMs for action recognition by Yamato et al. [144], uses silhouette images quantized into super-pixels, each pixel counting the ratio of black and white pixels within its underlying region, as features. A similar representation is also used in [130], see Figure 4 a) and b). In [7] silhouettes are integrated over time in so called *motion history images* (MHI) and *motion energy images* (MEI), see Figure 6 a). Instead of integrating a time sequence into a single image, [5, 147] work directly on the space-time volume spanned by a silhouette sequence over time, see Figure 6 c). Other silhouettes and contour based based representation have been used for instance in [85, 98, 17], see Figure 4 c).

One way to deal with noisy silhouettes, e.g. in outdoor scenes where exact background segmentation is difficult, is to use the *chamfer distance* as for instance in [26, 135], or by using *shape context descriptors* [70, 119, 150].

As demonstrated by many of the above mentioned approaches, silhouettes provide strong cues for action recognition, and moreover have the advantages of being insensitive to color, texture, and contrast changes. On the downside, silhouette base representations fail in detecting self-occlusions, and depend on a robust background segmentation.



(a) (b) (c) (d) (e)

Figure 4: Global posture representations: (a) Silhouettes of tennis strokes [144]; (b) silhouettes pixels accumulated in regular grid [130]; (c) spline contours [98]; (d) optical flow magnitude accumulated in regular grid [94]; (e) optical flow split into directional components, then blurred [25]. Awaiting publisher permission

A second important class of image models uses dense optical flow extracted from consecutive images. An early example of using optical flow for action recognition is given by Polana and Nelson [93], where they compute *temporal-textures*, i.e. first and second order statistics based on the direction and magnitude of normal flow, to recognize events such as motion of trees in wind or turbulent motion of water. In [94] Polana and Nelson propose features for human action recognition based on flow magnitudes accumulated in a regular grid of non-overlapping bins, see Figure 4 d). Another early approach which uses optical flow is proposed by Cutler and Turk [20], where the optical flow field is clustered into a set of *motion blobs*, and motion, size, and position of those blobs are used as features for action recognition.

More recently, [25] split the optical flow field into four different scalar fields (corresponding to the negative and positive, horizontal and vertical component of the flow), see Figure 4 e), which are separately matched. This representation was also used in [99, 134].

In the works [54, 61, 29], the adaboost-based Viola–Jones face detector is extended to action recognition by replacing the rectangular image features with spatio-temporal cubes computed over optical flow.

Flow based representations do not depend on background subtraction, which makes them more practical than silhouettes in many settings, because they do not require background models. On the downside, they rely on the assumption that image differences can be explained as a result of movement, rather than changes in material properties, lighting, etc.

Another important class of image features is based on gradients. [149] compute gradient fields in XYT direction and represent each frame through the histogram over those gradients. Also the HOG descriptor, which has been very successfully applied to person and object detection [22], has been used for action recognition [124]. Instead of computing a single gradient histogram per frame, the HOG descriptor divides the image grid into regular spaced overlapping blocks, and computes a histogram within each of those blocks.

Gradient features share many properties with optical flow features: they do not depend on background subtraction, but likewise are sensitive to material properties, textures, and lighting, etc. In contrast to optical flow, gradients are discriminative for both moving and non-moving parts, which has advantages as well as disadvantages. For instance static non moving body parts can also provide important cues for an action, but might be easily confused with static object in the background with strong gradients.

Recently several works demonstrated superior results by combining gradients and flows [61, 60], or silhouettes and flow [126].

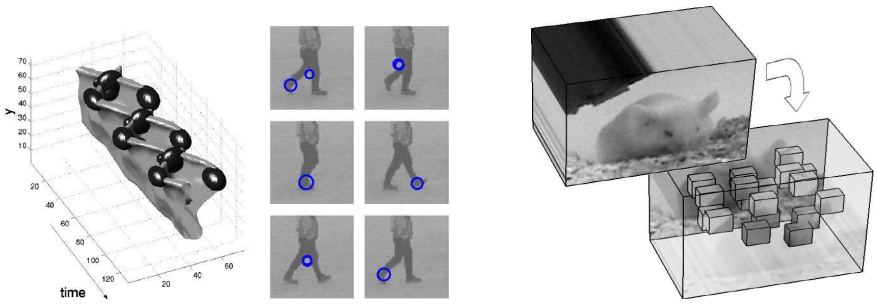
The last class of image models which we discuss, is based on the neuroscientifically inspired HMAX approach [113]. For instance the approaches [49, 109] combine Gabor filters and optical flow in a max-pooling scheme, to simulate the basic stimulus-response functions of a virtual cortex. Because low-frequency Gabor filters can have very similar shapes than oriented gradient filters, both provide similar cues. Yet by using higher-order Gabor filters, additional information can be introduced, which however also leads to a strong increase in computational time.

As explained earlier, image models of actions result in strong simplifications compared to parametric body models. One important consequence is that they are very sensitive to variations in the view direction of the camera and body sizes of the agent performing the action. It is thus important to account for such variation, either through a large number of different template instances, or by using suitable features and matching functions that are insensitive to such transformations.

Image-based representations have been used by many approaches of very different kinds, however, they are often based on strong assumptions that need to be addressed in future work. In particular many approaches assume that a ROI around a person, possibly even background subtracted, is provided by a previous processing stage. Consequently, these approaches strongly depend on the progress in related fields such as person detection and tracking. Also, most approaches only operate on fully visible bodies and do not investigate how to adapt global models to partial observations, e.g. occluded bodies or close-up views. Note anyway that video surveillance adapts well to these assumptions since far-views are frequent. Moreover, in such applications additional sensors, including time-of-flight cameras, sonars and tags, can alleviate poor background subtraction or poor motion analysis.

2.3. Spatial statistics

In this section, we review local representations of action which decompose the image/video into smaller regions, *not* linked to body parts or image coordinates. Instead, actions are recognized based on the statistics of local features from all regions. An immediate advantage of those approaches is that they neither rely on explicit body part labeling, nor on explicit human detection and localization.



(a) (b)

Figure 5: Local posture representations: (a) Space-time interest points in [59] are computed at points of high spatio-temporal variation ("spatio-temporal corners"). (b) Spatio-temporal features in [24] are designed to be more responsive than the former space-time interest points. Awaiting publisher permission

Local features can be computed in a dense or sparse set of regions. *Space-time interest points* [59, 24] were specifically introduced to generalize interest points and local descriptors [19, 68] already used in object recognition and image classification to the case of action recognition and video classification. Such approaches are typically based on bottom up strategies, which first detect interest points in the image, mostly at corner or blob like structures, and then assign each region to a set of preselected vocabulary-features. Image classification reduces then to computations on so called *bag of features* (BOF), i.e. histograms that count the occurrence of the vocabulary-features within an image. Similar interest point detectors for action recognition have been proposed by [59] and later by [24], see Figure 5.

The work [59] originally extend Harris corner detection [41] and automatic scale selection [63] to 3D space and time. The vocabulary features used in this work are the responses to a set of point-centered and scale-adapted higher order gradient filters. This work was extended to BOF and SVM classification in [110]. Dollar et al. [24] proposed an alternative interest point detector based on a quadrature pair of 1D Gabor filters applied temporally and spatially. This work also introduces several SIFT-like [68] space-time descriptors based on local PCA and histogramming of gradient, flow, or brightness values. Another interest point detector is proposed in [141], where an image sequence is decomposed into spatial components and motion components using non-negative matrix factorization (NMF). Interest points are then independently detected in 2D spatial and 1D motion space

using difference of Gaussian (DoG) detectors.

A practical advantage of interest-point approaches is that the detection of the agent need not be performed explicitly for the computation of the space-time features. The detected interest points need to show some consistency for similar observations, but usually they can also account for some outliers. On the downside, the detected features are usually unordered and of variable size, and consequently modeling geometrical and temporal structure is difficult with space-time features. Many approaches stick therefore with the previously mentioned *spatial bags of features* representation, which describes sequences simply through histograms of feature occurrences, hence without modeling any geometrical structure between the feature locations.

To add structural information, some approaches [82, 142, 31] use graphical models with hidden variables for the position of patches. In [34] so called *compound features* are proposed, which can be seen as some kind of super features taking into account the relative positions of several features in a neighborhood. Another possibility to add structural information is to divide the image space into several local BOF histograms [60]. Other interesting issues with BOF based approaches are: how to select a small but discriminative vocabulary [67], and how to combine different types of features, e.g. local features and silhouette based features [65].

Finally it is also important to mention, that although most of the previous approaches compute SIFT-like histograms over cubes in 3D space and time, the gradients used in the histograms are nevertheless mostly only 2D spatial. In fact, finding a uniform quantization for vectors on a 3D sphere is a well known problem, which was recently addressed by several papers [111, 57] in the context of deriving 3D SIFT descriptors for action recognition.

Though the majority of local feature representations is based on the previously discussed extensions of SIFT, several other local representations have been proposed. In [55] local patches are computed from a color-based over-segmentation of the space-time volume. Loosely spatial relations between the resulting segments are then learned via pictorial structures [32], and used for matching actions. The work [9] does not identify patches via segmentation or feature detectors, but searches instead over all possible images patch configurations of a given size.

In summary, statistical methods based on local features have recently drawn a lot of attention in the action recognition community because they promise the same advantages as in static object recognition, and because they can easily apply to difficult scenes, e.g. movies or video clips from the inter-

net, that evidently will be very difficult to model with full-fledged image or body models. However, the very nature of complex human actions will probably make it necessary to combine those methods with stronger spatial and temporal models, e.g. computing spatial statistics over dense (rather than sparse) image grids, and relying on human detection for scenes containing multiple persons. The combination of spatial statistics with strong temporal models (including grammars, templates and keyframes) will be further investigated in the next section.

3. Temporal Action Representations

In the previous section we discussed the different kind of image features that can be extracted from a video sequence to represent the spatial structure of actions. We will now describe the different representations that can be used to learn the *temporal structure* of actions from such features. As a result, we further classify approaches to action recognition, based on how they express the temporal component of the observations. We distinguish between three main categories of representations: grammars, templates and temporal statistics.

3.1. Action grammars

The approaches discussed in this section represent an action as a sequence of moments, each with their own appearance and dynamics. A common way to approximate a dynamical system over feature observations is to group features into similar configurations, i.e. states, and to learn temporal transition functions between these states. Such models fall generally into the class of *graphical models*, which are best described as probabilistic grammars.

Among the versatile probabilistic grammars used for action recognition the most prominent is certainly the *hidden Markov model* (HMM) [95]. The HMM came in particular to fame because of its great success in the speech and natural language processing community.

The first work on action recognition using HMMs is probably that of Yamato et al. [144], where a discrete HMM is used to represent sequences over a set of vector quantized silhouette features of tennis footage, see Figure 4 a). [121] use a continuous HMM for recognition of American sign language. [139] are recognizing hand gestures using a HMM. [13] learns a kind of *switching-state HMM* over a set of autoregressive models, each approximating linear motions of blobs in the video frame.

There are many other approaches using HMMs, to name a few: [10] investigated how a HMM can be learned in one space, e.g. parametric body poses, and mapped to another, e.g. 2D silhouette observation. [132] propose a distance measure between HMMs for unsupervised clustering of gestures. [69] use HMMs as weak classifiers in an adaboost based action recognition approach.

HMMs are purely sequential models of action, which severely limits their use for full-body action recognition, where the different body parts may move independently and in parallel. Various extensions to the more general class of *dynamic Bayesian networks* (DBN) have been proposed to overcome this limitation. [12] learn coupled HMMs to model interactions between several state variables, e.g. interactions between left and right hand motions. [87] use a complex DBN to model interactions between two persons.[91] model interactions between people and objects. [70, 136] extend HMMs with explicit latent states for view point, to model actions seen from arbitrary views in a single model. [76] use a mixture of Markov chains to model distributions of dense KLT trajectories.

A less obvious limitation of HMMs is that they are *generative models* of actions, which rely on simplifying statistical assumptions for computing the joint probability of the states and the observed features, whereas a more general *discriminative* model may better predict the conditional probability of the states *given* the observed features. As a result, several authors have investigated the use of discriminative models of actions.

[119] proposes to use Conditional Random Fields (CRF) instead of HMMs. CRFs are discriminative Markov models which can use non-independent features and observations overs time (contrary to the HMM assumption). Furthermore, CRF parameters can be trained to maximize the discriminative power of the classifier, rather than the joint probability of the training examples. Modeling sub-structures within actions is, however, *not* as straightforward with a CRF as with a HMM. This issue was overcome in [131, 78, 130] by using hierarchical layers of latent variables.

Other dynamic models that have been used for action recognition are: auto regressive models [13, 98, 4], time delayed neural networks [146], context-free grammars [48] and feature-structure grammars [58, 124].

A strong advantage of action grammars is their high degree of modularity. This makes them suitable for generalizing over large variations in acting speeds and styles. Grammars are also compositional, i.e. grammar models of primitive actions can also serve as smaller vocabulary units to build larger

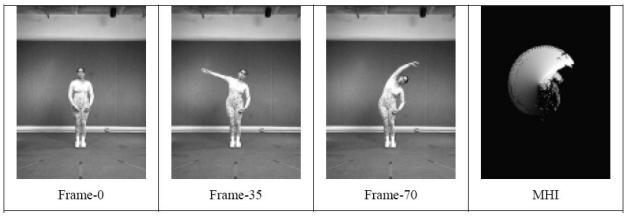
networks of complex actions [45], and similarly, complex models can be used to segment sequences into smaller units [11, 36, 89], as discussed more in detail in Section 4.3. Parameters of probabilistic grammars can also be learned quite efficiently, using small numbers of labeled examples in supervised mode, or large numbers of non-labeled examples in non-supervised mode. But the structure of probabilistic grammars must usually be chosen manually (see Kitani et al. [56] for an notable exception). As a result, learning and evaluation of grammar-based action recognition remains an outstanding problem with large numbers of actions classes.

3.2. Action templates

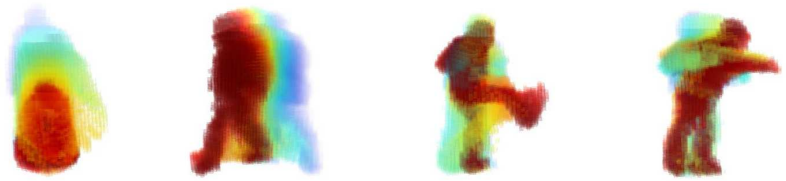
Instead of representing features and dynamics explicitly and separately in a layered model, some methods attempt to directly learn the appearance of complete temporal blocks of features - which we call templates. Typically, template-based approaches directly represent dynamics through example sequences, either by stacking features from several frames into a single feature vector, or by extracting features from the n-dimensional *space-time volume* spanned by a sequence over time, see also Figure 6. Though most of the approaches that use templates are based on image models, such as [5, 147] that build templates by stacking multiple silhouette images into a single volumetric representation, they can also be used with parametric models [40, 83, 143, 97, 37] or even local representations [55].

Templates are typically computed over long sequences of frames, and should not be confused with spatio-temporal features or optical flow (Section 2.2 and 2.3), which are computed over small time windows (typically 2-4 frames) and serve as components of other action classifiers. The seminal work on *action templates* is that of Bobick and Davis [7], who build a motion history image (MHI) by mapping successive frames of silhouette sequences into a single image, see Figure 6 a). MHIs are generally similar to a depth map computed from a space-time volume. Various variants of MHIs have been proposed in [74, 75, 14], see Figure 6 b). MHIs have been also extended to motion history volumes (MHVs) [138], see Figure 6 b), by using visual hulls computed from multi view sequences instead of the 2D silhouettes used in the original work.

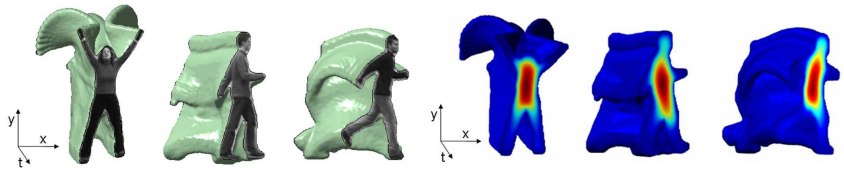
Templates are usually fixed-size vector representations, which makes them straightforward to implement in combination with most static classification techniques. Often simple nearest-neighbor assignment or naive Bayes classification are used in experiments. Contrary to grammars and state-transition



(a)



(b)



(c)

Figure 6: Action template methods: (a) motion history images (MHI) [7]; (b) motion history volumes (MHV) [138]; (c) space-time shapes [5]. Awaiting publisher permission

models, templates *cannot* represent variations in time, speed, and action style through special variables. Variations are instead implicitly represented through large sets of example sequences, making the classification problem more difficult. In those cases, advanced statistical learning methods have been proposed, such as neural networks [40], support vector machines [75] and adaboost [61, 54, 29].

To deal with actions with variable durations, an additional normalization step may be necessary to ensure that the resulting feature vectors have the same dimension, or the more advanced dynamic time warping (DTW) [23, 83, 33, 128] may be used.

Instead of using multiple templates in a conventional classifier, [100] proposes to build a single template from a collection of templates using a MACH filter, which can then simply be correlated with new observations sequences.

Other important examples of action templates use Fourier or wavelet representations in the temporal domain, [93, 64]. Trajectories of body parts or image features can be also used as templates. For instance, [71] introduce templates of body feature trajectories after tracking over extended time sequences.

In summary, template based representations of very different kind have been proposed. Generally they are effective and discriminative action representations, and in particular attractive for action classification tasks because they straightforward integrate with powerful static classifiers such as SVMs or adaboost. On the down side, they are less amenable to action detection tasks because they do not have efficient methods for temporal segmentation (see Sections 4.3 and 4.2 for a comparison between template-based and grammar-based action segmentation), and they do not generalize very well to incomplete or missing observations, e.g. occlusions.

3.3. Temporal statistics

In contrast to previous references, some approaches attempt to build statistical models of the appearance of actions, without an explicit model of their dynamics. Typical examples are methods that learn an appearance model of action from a single characteristic *key-frame* as in a photograph [133, 62] or from the histograms of (image, body or local) features over time.

Carlsson and Sullivan [17] introduce the use of *key-frames*, i.e. a single characteristic frames of an action, to recognize forehand and backhand strokes in tennis recordings. Matching in [17] is based on a sophisticated

point to point matching between edge filtered images, to measure the deformation of a edge template with respect to the image observation. [109] extend key-frames to very short *snippets* of frames, raising the interesting issue of how many frames are required to perform action recognition.

Besides single static images, sequences can be also encoded without taking temporal relations into account. Histogram techniques, i.e. *temporal bag of features* approaches, have been used to represent sequences simply base on the frequency of feature occurrence over time [110, 24, 81, 111, 134]. The biologically motivated system of [49] uses a different technique with feature vectors computed as maximum match responses to a set of prototypes. Similarly, [135] use an *exemplar-based embedding*, which represents a sequences via its minimum distances to a set of prototypes.

An extension to temporal BOF based on "temporal binning" is proposed in [84] to explicitly take into account the short-time temporal ordering of visual words. *Spatial-temporal correlograms* have also been proposed for adding temporal and spatial dependencies to BOF methods [111, 67, 108]. [107] model explicit temporal (before and after) and (near and far) spatial relations between spatio-temporal interest points. In [60] a sequence is split into several localized BOF histograms to add temporal and spatial constraints.

Clearly, such *dynamic-free* representations cannot be applied to discriminate all kind of actions, e.g. two actions that share similar poses but in different temporal order. However, there is growing interest in such representations, and in particular the BOF approach. This is partially imposed by the recent trend in using local features, which simply seem to work best with such a simplified representation. Nonetheless, and in particular for modeling small atomic motions, they have powerful properties, such as being efficient to compute, insensitive to timescale variation, and very discriminative, which makes them very attractive for large-scale action recognition.

4. Action Segmentation

In the first two sections of this survey, we have mostly been concerned with approaches that extract visual features from video streams and combine them in space and in time for making a decision on what actions are present in the video. In many cases, those approaches are demonstrated with results obtained using *segmented* video clips each showing a single action from start to finish, both for training and testing. But can the two tasks of *action segmentation* and *action recognition* really be performed separately?

There appears to be very little evidence from neuroscience on how motion segmentation and recognition interact in the human visual system. From a computational point of view, it is of course beneficial to segment the video stream before applying recognition, since labeling a segment is much more efficient than labeling all subsegments in a stream. But this raises the difficult issue of finding a generic vocabulary of *parts of actions*, and generic methods for breaking video streams into the corresponding segments. In practice, this appears to be a problem no less difficult than action recognition itself. This section discusses different methods used for temporal segmentation. We classify those methods into three broad classes: boundary detection, sliding windows and grammar concatenation.

4.1. Boundary Detection

A common strategy for recognizing actions is to use a generic segmentation method based on detecting motion boundaries, then separately classify the resulting segments. Such motion boundaries are typically defined as discontinuities and extrema in acceleration, velocity, or curvature of the observed motions. The choice of boundaries thus implicitly results in a basic motion taxonomy.

An early paper by Marr and Vaina [73] discusses the problem of segmenting the 3D movement of the human body, and suggests the use of rest states, i.e. local minima, of the 3D motion of the limbs as natural transitions between primitive movements. Similar, Rubin and Richards define in their work [105] two elementary kinds of motion boundaries: *starts and stops* and *dynamic boundaries*. *Starts and stops* are analog to the rest states defined by [73]. *Dynamic boundaries* appear between starts and stops and result from discontinuities, e.g. steps or impulses, in force applied to the object in action.

Following this theoretical line of research, computational approaches for motion boundary detection have been proposed. [106] perform an SVD decomposition of a long sequence of optical flow images and detect discontinuities in the trajectories of selected SVD components to segment video into motion patterns. Also [85] segment action sequences by detecting minima and maxima of optical flow inside body silhouettes. In [129] impulses in motion, so called *ballistic dynamics*, are used to detect motion boundaries. Other approach to detect motion boundaries are: [137] that computes motion features based on visual hulls, [132, 97] that uses 2D trajectories of hands,

see Figure 7(b), and [52] that uses a hierarchical body model. In [14] motion boundaries are detected using sequential change detection methods.

In theory, boundary detection methods are attractive because they provide a generic segmentation of the video, which is not dependent on the action classes. In practice, the segmentation must be used with some precautions because (a) they are subject to errors in the recovery of the motion field; (b) they are not stable across view-points; and (c) they are easily confused by the presence of multiple, simultaneous movements.

4.2. Sliding Windows

Another strategy for recognizing actions divides the video sequence into multiple, overlapping segments, using a sliding window. Classification is performed sequentially on all the candidate segments, and peaks in the resulting classification scores are interpreted as action locations. In contrast to boundary detection methods, the segmentation here depends very strongly on the recognition stage. As a result, it should be clear that those methods are not applicable in the training stage. A consistent segmentation of the training examples must be provided manually or through another method, and is a crucial elements for the success or failures of those methods.

A sliding window approach can be used with any of the previously discussed feature representations and classifiers. Many template-based representations [149, 153, 30, 54, 55] use a sliding window. Some approaches use them in combination with dynamic time warping (DTW) [23, 79] and even grammars [6, 140].

Compared to boundary detection methods, sliding window methods are usually much more computationally intensive, as they involve many evaluation of all classifiers. To achieve robustness against the duration of actions, they often require multiple window sizes as well, which results in an additional computational burden. Sliding window methods may also produce unpredictable results in the presence of unknown action categories. However, sliding window methods make less assumptions, i.e. they do not assume special boundary criteria, and can be easily integrated on top of any action classifier without requiring further computation of special segmentation features.

4.3. Grammar concatenation

In Section 3.1 we reviewed representations of individual action classes with grammars, which give a model of the transitions between states in the

action. This suggests a general approach for segmenting actions by *concatenating* action grammars to model the transitions *between* actions as well. Indeed, this provides an effective means of simultaneously segmenting and recognizing actions. Concatenative grammars can be build for instance by joining all models in a common start and end node and by adding a loop-back transition between these two nodes. It is also possible to allow for more complex transitions between actions, e.g. actions may share states and transitions between actions may be adjusted individually to reflect more realistic the probabilities of one actions following another. Such complex structure are similar to HMM networks used in continuous speech recognition. Segmentation and labeling of a complex action sequence is then computed as a minimum-cost path trough the network using dynamic programming techniques, e.g. the Viterbi path for HMMs [95]. The works [11, 36, 89, 69] use such networks for action recognition based on HMMs. Similar [119, 78] use CRFs, and [116] Semi-Markov models. The work of [98] uses autoregressive models to represent actions, and a condensation filter to switch between these models.

Those approaches make neither the assumptions of the boundary detection methods, nor do they require heavy evaluations such as sliding window approaches. The segmentation is elegantly and efficiently solved using dynamic programming techniques. However, it should be emphasized that learning a concatenative grammar with many actions requires a much larger amount of training data, especially when transitions between actions are learned from real data. In speech recognition such data is available in form of text-documents, word-transcriptions, and phonetically labeled sequences. Similar data does, however, currently not exist for action recognition, and therefore transitions between actions are often set manually, or with strong assumptions, such as uniform transition probabilities.

5. View-Independent Action Recognition

As mentioned earlier (Section 2.1), fundamental considerations on the model representation, i.e. whether to use a 2D or 3D representation, have a long history in action recognition, and approaches demonstrating the qualities of either direction have been proposed.

Following the initial success of those approaches, new challenges, such as learning larger number of action classes and robustness under more realistic settings, gained importance. Within this scope a very important demand

is independence to viewpoint, which wasn't address by most of the early approaches. It is our opinion, that such considerations bring the issue on how to represent posture, i.e. in 2D or 3D, into a interesting new perspective.

We take our taxonomy for view-independent action recognition from work on view-independent shape matching [53], which names three strategies for view-independent matching: normalization, invariance, and exhaustive search. In the following we discuss approaches based on these strategies, and separate as well between view representations in 2D or 3D.

5.1. View normalization

During view-normalization, each observation is mapped to a common canonical coordinate frame. Therefore normalization approaches generally first estimate cues that indicate the transformation from the canonical view frame to the current view of the observation, and then correct the observation with respect to the estimated transformation. Matching then takes place in the normalized coordinate frame.

5.1.1. Normalization in 2D

Normalization is used by many approaches as a preprocessing step to remove global scale and translation variations. In particular image models (Section 2.2) often extract a rectangular ROI around the subject, and scale and translate this region to a unit frame. This normalization removes global variations in body size, as well as some scale and translation variations resulting from perspective changes.

Normalization with respect to out-of-plane transformations, e.g. a camera rotation, is not trivial given a single 2D observation. Nevertheless, [101] propose a method, which estimates the 3D orientation of a person from its walking direction in 2D, using knowledge about the ground homography and camera calibration. Assuming only horizontal rotation of the body in 3D, the 2D silhouette of the person is perspectively corrected onto a fronto parallel view and matched against a set of canonical silhouettes.

5.1.2. Normalization in 3D

Although it somehow limits the application of action recognition approaches, walking direction as orientation cue was as well used by several 3D based approaches to compute a reference frame for normalization [8, 21, 102, 151, 90]. Given a 3D body model, an orientation independent joint representation can be computed based on the global body orientation.

Often the torso is used as reference object to normalize all joints with respect to its orientation. It is further possible to represent each body part with and individual coordinate frames. For example, [33] compute individual reference frames for the torso, arms, and hips.

In summary, normalization approaches are based on the estimation of the body orientation. If strong cues, such as walking direction or a reconstructed body model are available, the orientation can be easily derived. However, all following phases depend on the robustness of this step. Misalignments, because of noisy estimations or intraclass variations, are likely to affect all following phases of the approach.

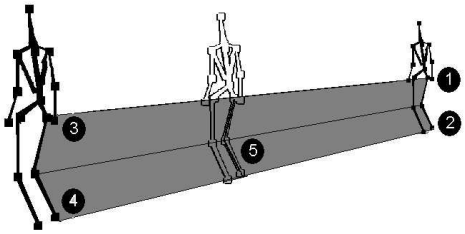
5.2. View Invariance

View-invariant approaches do not attempt to estimate view transformations between model and observation. Instead view-invariant approaches search for features and matching functions that are independent with respect to the class of view transformations considered.

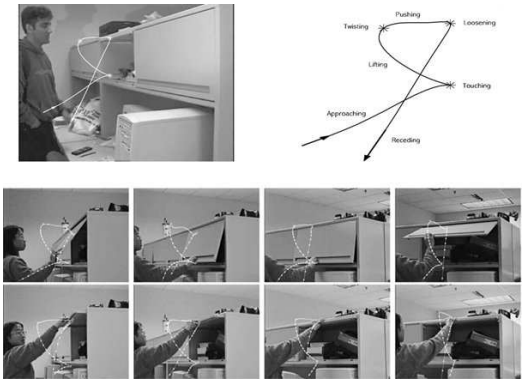
5.2.1. View Invariance in 2D

A simple form of view-invariance is based on histogramming. Instead of representing image features in a fixed grid, only the frequency of feature occurrences is stored. Such an representation has been used for instance by [149] to represent distributions of space-time gradients. This representation, however, only provides invariance to translations in the image plane.

The availability of point correspondences, e.g. in form of anatomical landmarks, was frequently used for view-invariant matching between pairs of observations, see Figure 7 for some examples. For instance, an epipolar geometry can be estimated from a subset of point correspondences and then used to constrain the set of all point correspondences, and respectively a matching cost over changing views can be computed without requiring a full 3D reconstruction. I.e. given point matches (x_i, x'_i), $i = 1, \ldots, n \geq 8$ in pairs of images I, I', the fundamental matrix F, which holds the relation $x_i F x'_i = 0$, can be estimated. This relation holds however only if all point pairs come from the same rigid object. Hence the resulting residual $\sum_i |x_i F x'_i|^2$ can be used as matching cost [123, 37, 114, 147, 148, 115]. Similar, matrix factorization and rank constraints, as in structure from motion estimation [125], can be used to validate whether point correspondences in two images came from the same single rigid object [112, 97].



(a)



(b)

Figure 7: View invariant action recognition: (a) geometrical invariants can be computed from 5 points that lie in a plane [86]; (b) View-invariant matching of hand trajectories [97]. Point matches between different observations are computed from discontinuities in motion trajectories. Awaiting publisher permission

Geometric invariants, i.e. measures that do not change under a geometric transformation, can also be used for invariant matching of landmark points. These invariants can be computed from 5 points that lie in a plane [86].

More recently, an invariant approach that optionally uses point correspondences or image features is proposed in [51], based on frame-to-frame self-similarities within a sequence. The representation discards all information related to an absolute reference frame and is only based on the relative change between frames. It is shown in [51] that such features remain surprisingly stable under changing viewing conditions. Farhadi et al. [28] provides some kind of view-invariance by using a *transfer learning* approach, which maps an action model from a *source-view* into a novel *target-view*. To establish such a transfer mapping, explicit samples of corresponding observations from source and target view must be available during learning, those need however not provide views of the same action class, for which the transfer function is learned.

5.2.2. View Invariance in 3D

Campbell et al. [15] investigate 10 different view-invariant representations based on 3D body part trajectories. These include shift invariant velocities (dx, dy, dz) in cartesian coordinates, and shift and horizontal rotation invariant velocities $(dr, d\theta, dz)$ in polar coordinate. In evaluation on 18 Tai Chi gestures, the polar coordinate representation has best overall recognition rates. [18] proposes a view invariant pose representation based on a voxel reconstruction and cylindrical 3D histograms similar to the 2D shape context descriptor [3]. The same descriptor was later used by [92] for action recognition. Another invariant representation based on 3D shape-context and spherical harmonics was proposed in [43]. [138] proposes a view-invariant representation in 3D, based on Fourier coefficients in cylindrical coordinates. The same representation was later used in [127], but with a more sophisticated modeling approach based on Stiefel and Grassmann manifolds.

In summary, invariant approaches remove view-dependent information during feature computation. Computing a single view-independent feature per action is certainly more efficient than considering all possible views, and moreover does not depend on a possible false recovery of view orientation. However, removing view-dependent information will usually always also result in a loss of discriminative information.

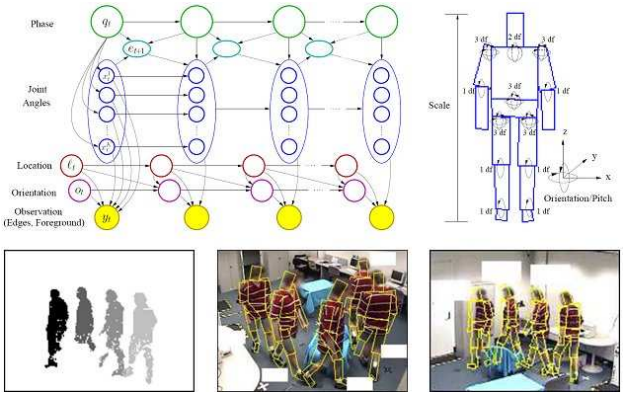


Figure 8: Generative MOCAP: Hierarchical HMM body model and tracking results using the generative approach in [90]. Awaiting publisher permission

5.3. Exhaustive Search

Instead of deciding on a single transformation, as it is typical for normalization methods, or discarding all transformation dependent information, as with invariant methods, one can search over all possible transformations considered.

5.3.1. Exhaustive Search using Multiple 2D Views

Several approaches use a fixed set of cameras installed around the actor, and simultaneously record the actions from this multiple views. During recognition, an observation is then matched against each recorded view and the best matching pair is identified. In their work on MHIs, [7] record actions with 7 cameras, each with an offset of 30° in the horizontal plane around the actors. During recognition two cameras with 90° offset are used, and matched against all pairs of recorded views with the same 90° offset. Similar [85, 2] use 8 prerecorded views, and a single view during recognition.

5.3.2. Exhaustive Search using a 3D Model

To achieve more flexibility with respect to changes in camera setup, an internal model based on a 3D representation can be used. From such a 3D representation, and given camera parameters, any possible 2D view observation can be rendered. Such *generative approaches* are frequently used in MOCAP, where parameterized 3D models of the human body are projected into 2D. These models have explicit variables for global 3D position

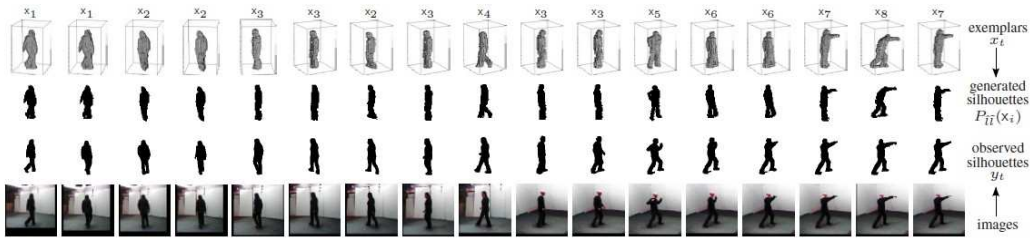


Figure 9: Generative model based on set of exemplary 3D key-poses. Used for view-independent action recognition in [136]. Awaiting publisher permission

and orientation, that are estimated simultaneously with the remaining joint parameters, e.g. [117, 90], see also Figure 8.

Similar methods haven been proposed for action recognition, and extended such that they do not require a joint model. The approach [70] uses a small set of synthetic 3D key-poses, rendered from a modeling software. Observations are then compared against the silhouettes, produced by projecting the poses into 2D with respect to all possible view transformations. Dynamics over poses and changes in view transformations are modeled in a dynamic network, and the best pose-view sequence is found via a dynamic programming search. The approach [136], see Figure 9, shares as well the idea of projecting a set of learned 3D key-poses into 2D to infer actions from arbitrary views. This work uses a HMMs with additional capabilities to model unknown view transformations. Another approach based on the same ideas is proposed in [80], which uses a CRF instead of the HMM.

Instead of using a data-driven approach to producing the 2D projections, [120] directly learn an analytical function, which takes as input the viewpoint in 3D and outputs the corresponding silhouette representation in 2D. Another direction is taken in [145]. Instead of projecting a 3D model into 2D, features are detected first in 2D, and then back-projected onto *4D action shapes*. This approach requires as well an optimization over the possible view orientations to find the best 2D-3D matching.

In summary, approaches based on exhaustive search have recently gained some interest, partially because their computational expense is about to become fairly manageable with modern computer systems. They neither depend on deterministic detections of body orientation, nor do they discard discriminative information during an invariant feature computation step. They, however, depend strongly on the availability of specially recorded and an-

notated multi-view datasets, and usually require assumptions on the search space, i.e. restrictions to certain classes of view-transformations.

6. Dataset

Finally, we want to discuss some of the dataset, which are currently used by many of the action recognition approaches as a benchmark. Unfortunately acquiring realistic action footage, including ground truth data, is a very difficult and time consuming task, and currently there is certainly lack of such data in action recognition.

The three popular datasets, which are currently used by most of the approaches are: *KTH* [110], *Weizmann* [5], and *IXMAS* [138]. They all contain around 6-11 action performed by various actors. They are all not very realistic and share strong simplifying assumptions, such as static background, no occlusions, given temporal segmentation, and only a single actor.

The recognition rates of the papers discussed in this survey on those datasets are given in Table 2. It is however important to note that not all approaches follow the exactly same evaluation methodologies, so approaches can't be compared purely based on those results. Moreover, in light of the simplifying assumptions made in the datasets, it is not evident how those results might extrapolated to more complex scenarios. The three datasets are detailed in the following.

6.1. The KTH Dataset

The KTH dataset [110], Figure 10, contains the six actions *walking, jogging, running, boxing, hand waving and hand clapping*, performed several times by 25 subjects in four different scenarios. Overall it contains 2391 sequences. It has fewer action classes than the two other datasets, but the most samples per class. It is hence well suited for learning intensive approaches, e.g. approaches based on SVMs.

In difference to the two other datasets it does not provide background models and extracted silhouettes, and moreover some of the scenes are recorded with a shaking and zooming camera. Most approaches that evaluate on the KTH dataset are hence based on local features (Section 2.3) which are best suited to such scenarios. Recently, however some approaches [49, 109] that require person detection reported as well results. The original paper [110] reported a recognition rate of 71.7% on the dataset. More recently several approaches reported recognition rates above 90% up to 94%.

Figure 10: Example images from the KTH dataset

The results of the different approaches are shown in Table 2. Note however, as also pointed out in [60], not all approaches follow the same evaluation methodology of the original paper, which makes a direct comparison difficult. In the original paper the data was split into a training set (8 persons), a validation set (8 persons), and a test set (9 persons). In the table we distinguish between approaches that use this split, and those that use a leave-one-out cross-validation. Note that the latter usually gives better results, because more data is available for training.

6.2. The Weizmann dataset

The Weizmann dataset [5], Figure 11, contains the nine actions *running, walking, bending, jumping-jack, jumping-forward-on-two-legs, jumping-in-place-on-two-legs, galloping-sideways, waving-two-hands, waving-one-hand*, performed by nine different actors. Overall it contains 93 sequences, all performed in front of similar plain backgrounds, and with a static camera. It is the smallest of the three datasets considered.

The original approach [5] reported already a very high recognition rate of 99.6%, and similar results have been archived by many of the subsequent approaches. It appears hence to be the easiest of the three datasets. Nevertheless it is still used in recent works.

Generally approaches which use the background subtracted silhouettes achieve best rates (up to 100%). Recently also several approaches that only depend on person location, but not on extracted silhouettes could report very high recognition rates, e.g. [109, 135].

The reported average recognition rates of the different approaches are shown in Table 2. Note however that approaches are using slightly different evaluation methodologies, which makes direct comparison purely based on these values difficult. For instance, some of the approaches use only eight

Figure 11: Example images from the Weizmann dataset

of the nine actions, some evaluate on small segments others on the complete sequences, etc.

6.3. The IXMAS dataset

The INRIA XMAS dataset [138], Figure 12, contains the 11 daily-life actions: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *pick-up*, performed each 3 times by 11 non-professional actors. Note that there are two more actors and actions on the dataset's website, but those have not been used by most of the approaches. The actions were filmed with 5 carefully calibrated and synchronized cameras. Overall it contains hence 429 multi-view sequences, or, if the views are considered individually, 2145 sequences. It also provides background subtracted silhouettes and reconstructed visual hulls.

The scenes are recorded in front of simple static studio-like backgrounds. Its main difficulty comes from the changing viewpoint, that is caused by the different camera configurations and the fact that actors freely chose their orientation while performing the actions. Respectively, the dataset is in particular used by view-independent approaches (Section 5).

The best known recognition rates were recently reported by [127] (98.78%) using 3D MHVs [138] and a modeling approach based on Stiefel and Grassmann Manifolds. Approaches which use only a single camera for recognition reported results up to 82%.

Results of the different approaches are shown in Table 2, and we distinguish between approaches working in 2D and in 3D. Moreover, and similar as for the two other datasets, the evaluation methodologies of the different approaches vary slightly, which makes direct comparison difficult. For instance, some of the approaches use only a subset of the provided sequences for evaluation.

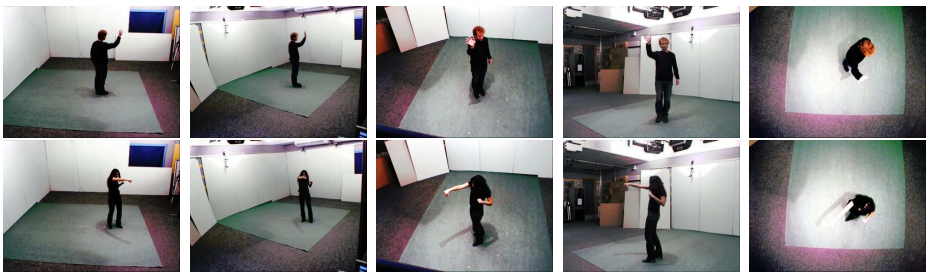


Figure 12: Example images from the five view used in the IXMAS dataset

6.4. Other datasets

There are several other not so frequently used datasets available. The CMU MoBo database [38] and the HUMAN-EVA database [118] were primarily designed for motion capture and hence only contain very simple action. Nevertheless they have been used by some approaches in action recognition. The HOHA datasets [60] are a large collection of short segments of real *Hollywood* movies, annotated with 12 action classes: *answer-phone, drive-car, eat, fight-person, get-out-car, hand-shake, hug-person, kiss, run, sit-down, sit-up, stand-up.* The actions are performed by professional actors, under a wide range of camera viewpoints and in very different styles. This is a very challenging dataset, including inter-actions with people (fight-person, hand-shake, hug-person, kiss) and objects (answer-phone, drive-car, getout-car), which are outside the scope of this survey.

year	paper	spatial	temp	segm	view	KTH		Weiz	IXMAS	
						org	loo		2D	3D
1978	Marr [72]	bm	-	-	-	-	-	-	-	-
1982	Marr [73]	bm	-	bd	-	-	-	-	-	-
1985	Rubin [105]	-	-	bd	-	-	-	-	-	-
1989	Goddard [35]	bm	gr	bd	-	-	-	-	-	-
1992	Polana [93]	im	tmp	-	-	-	-	-	-	-
1992	Yamato [144]	im	gr	-	-	-	-	-	-	-
1993	Darrell [23]	im	tmp	-	-	-	-	-	-	-
1994	Guo [40]	bm	tmp	-	-	-	-	-	-	-
1994	Niyogi [83]	bm	tmp	-	-	-	-	-	-	-
1994	Polana [94]	im	tmp	pd	-	-	-	-	-	-
1995	Campbell [16]	bm	gr	sw	-	-	-	-	-	-
1995	Gavrila [33]	bm	tmp	-	norm	-	-	-	-	-
1996	Campbell [15]	bm	gr	-	inv	-	-	-	-	-
1997	Brand [12]	bm	gr	-	-	-	-	-	-	-
1997	Bregler [13]	bm	gr	sw	-	-	-	-	-	-
1997	Seitz [112]	bm	tmp	-	inv	-	-	-	-	-
1998	Bobick [6]	bm	gr	sw	-	-	-	-	-	-
1998	Yacoob [143]	bm	tmp	-	-	-	-	-	-	-
1999	Brand [10]	im	gr	gr	exh	-	-	-	-	-

1999	Rittscher [98]	im	gr	gr	-	-	-	-	-	-
2000	Brand [11]	bm	gr	gr	-	-	-	-	-	-
2000	Rui [106]	-	-	bd	-	-	-	-	-	-
2001	Bissacco [4]	bm	gr	-	-	-	-	-	-	-
2001	Bobick [7]	im	tmp	sw	exh	-	-	-	-	-
2001	Carlsson [17]	im	key	sw	-	-	-	-	-	-
2001	Syeda-Mahmood [123]	im	tmp	-	inv	-	-	-	-	-
2001	Wang [132]	bm	gr	bd	-	-	-	-	-	-
2001	Zelnik-Manor [149]	im	tmp	sw	-	-	-	-	-	-
2002	Kojima [58]	bm	gr	-	-	-	-	-	-	-
2002	Rao [97]	bm	tmp	bd	inv	-	-	-	-	-
2002	Zhao [151]	bm	gr	-	norm	-	-	-	-	-
2003	Bodor [8]	im	tmp	-	norm	-	-	-	-	-
2003	Cohen [18]	im	-	-	inv	-	-	-	-	-
2003	Efros [25]	im	key	sw	-	-	-	-	-	-
2003	Elgammal [26]	im	gr	-	-	-	-	-	-	-
2003	Kahol [52]	bm	-	bd	-	-	-	-	-	-
2003	Laptev [59]	ss	tmp	-	-	-	-	-	-	-
2003	Masoud [74]	im	tmp	-	-	-	-	-	-	-
2003	Parameswaran [86]	bm	tmp	-	inv	-	-	-	-	-
2003	Park [87]	bm	gr	-	-	-	-	-	-	-
2003	Ramanan [96]	bm	gr	-	exh	-	-	-	-	-
2003	Cuzzolin [21]	im	gr	-	norm	-	-	-	-	-
2004	Green [36]	bm	gr	gr	-	-	-	-	-	-
2004	Gritai [37]	bm	tmp	-	inv	-	-	-	-	-
2004	Ogale [85]	im	gr	-	exh	-	-	-	-	-
2004	Schuldt [110]	ss	ts	-	-	71.7	-	-	-	-
2004	Zhong [153]	im	tmp	sw	-	-	-	-	-	-
2005	Blank [5]	im	tmp	sw	-	-	-	99.6	-	-
2005	Boiman [9]	ss	key	sw	-	-	-	-	-	-
2005	Dollar [24]	ss	ts	-	-	-	81.2	-	-	-
2005	Feng [30]	im	tmp	sw	-	-	-	-	-	-
2005	Ke [54]	im	tmp	sw	-	63.0	-	-	-	-
2005	Peursum [91]	bm	gr	sw	-	-	-	-	-	-
2005	Robertson [99]	im	gr	sw	-	-	-	-	-	-
2005	Sheikh [114]	bm	tmp	-	inv	-	-	-	-	-
2005	Sminchisescu [119]	im	gr	gr	-	-	-	-	-	-
2005	Yilmaz [147]	im	tmp	-	inv	-	-	-	-	-
2005	Yilmaz [148]	bm	tmp	-	inv	-	-	-	-	-
2006	Ahmad [2]	im	gr	-	exh	-	-	-	-	-
2006	Kitani [56]	bm	gr	-	-	-	-	-	-	-
2006	Lv [69]	bm	gr/tmp	gr	-	-	-	-	-	-
2006	Niebles [81]	ss	ts	-	-	-	81.5	-	-	-
2006	Pierobon [92]	im	tmp	-	inv	-	-	-	-	-
2006	Rogez [101]	im	-	-	norm	-	-	-	-	-
2006	Roh [102]	im	tmp	-	norm	-	-	-	-	-
2006	Veeraraghavan [128]	im	tmp	-	-	-	-	-	-	-
2006	Wang [133]	bm	gr	sw	-	-	-	-	-	-
2006	Weinland [138]	im	tmp	bd	inv	-	-	-	-	93.3
2007	Guerra-Filho [39]	bm	gr	gr	exh	-	-	-	-	-
2007	Ikizler [46]	bm	bow	-	-	-	-	100	-	-
2007	Ikizler [45]	bm	gr	gr	exh	-	-	-	-	-
2007	Jhuang [49]	im	key	-	-	-	-	98.8	-	-
2007	Ke [55]	ss	tmp	sw	-	-	-	-	-	-
2007	Laptev [61]	im	tmp	sw	-	-	-	-	-	-
2007	Li [62]	ss	-	-	-	-	-	-	-	-
2007	Lv [70]	im	gr	-	exh	-	-	-	80.6	-

2007	Meng [75]	im	tmp	-	-	80.3	-	-	-	-
2007	Morency [78]	bm	gr	gr	-	-	-	-	-	-
2007	Niebles [82]	ss	ts	-	-	-	-	72.8	-	-
2007	Nowozin [84]	ss	ts	-	-	84.7	-	-	-	-
2007	Peursum [90]	bm	gr	-	norm	-	-	-	-	-
2007	Scovanner [111]	ss	ts	-	-	-	-	82.6	-	-
2007	Wang [130]	im	tmp	-	-	-	-	100	-	-
2007	Wang [134]	im	ts	-	-	-	92.4	-	-	-
2007	Weinland [136]	im	gr	-	exh	-	-	-	81.3	-
2007	Wong [141]	ss	ts	-	-	-	81.0	-	-	-
2008	Farhadi [28]	im	tmp	-	inv	-	-	-	58.1	-
2008	Fathi [29]	im	tmp	-	-	-	-	100	-	-
2008	Filipovych [31]	ss	ts	-	-	-	-	88.9	-	-
2008	Gilbert [34]	ss	ts	-	-	89.9	-	-	-	-
2008	Holte [43]	im	tmp	-	inv	-	-	-	-	-
2008	Junejo [51]	im/bm	tmp	-	inv	-	-	95.3	72.7	-
2008	Klaser [57]	ss	ts	-	-	91.4	-	84.3	-	-
2008	Laptev [60]	ss	ts	-	-	91.8	-	-	-	-
2008	Liu [67]	ss	ts	-	-	-	94.2	-	82.8	-
2008	Liu [65]	im/ss	ts	-	-	-	-	89.3	78.5	-
2008	Natarajan [80]	im	gr	-	exh	-	-	-	-	-
2008	Rodriguez [100]	im	tmp	-	-	-	-	-	-	-
2008	Schindler [109]	im	key	-	-	-	-	100	-	-
2008	Shen [115]	bm	tmp	-	inv	-	-	-	-	-
2008	Shi [116]	ss	gr	gr	-	-	-	-	-	-
2008	Souvenir [120]	im	tmp	-	exh	-	-	-	-	-
2008	Turaga [127]	im	gr	-	-	-	-	-	-	98.8
2008	Thurau [124]	im	ts	-	-	-	-	94.4	-	-
2008	Tran [126]	im	key	-	-	-	-	100	81	-
2008	Vitaladevuni [129]	im	gr	bd	-	-	-	-	-	87.0
2008	Weinland [135]	im	ts	-	-	-	-	100	-	-
2008	Yan [145]	im	tmp	-	exh	-	-	-	78.0	-
2008	Zhang [150]	im	ts	-	-	-	91.3	92.9	-	-
2009	Messing [76]	ss	gr	-	-	74.0	-	-	-	-
2009	Ryoo [107]	ss	ss	-	-	91.1	-	-	-	-

Table 2: We list the papers discussed in this survey with respect to the spatial representation used (spatial): body model (bm), image model (im), or spatial statistics (ss); the temporal model (temp): grammar (gr), templates (tmp), temporal statistics (ts), or temporal statistics (ts); the temporal segmentation (temp): boundary detection (bd), sliding window (sw), grammar (gr); and the view-independence representation (view): normalization (norm), invariance (inv), or exhaustive search (exh). We also show the recognition rates reported for the datasets: KTH, Weizmann, and IXMAS. For KTH we distinguish between approaches that use the data-split described in the original paper (org), and those that use a leave-one-out cross-validation (loo). We only show results for approaches that follow one of these two strategies.

7. Conclusion

In this paper we have given a survey of work in action recognition. We have classified approaches with respect to how they represent the spatial

and temporal structure of actions, how they segment and recognize actions from a continuous video stream, and how they handle variations in camera viewpoint. We identified a large body of different proposals and selected 150 representative papers. The survey reveals important progress made in the last ten years in small-vocabulary, single-person, full body action recognition. Important issues that must still be addressed in future work are scalability of action recognition systems with respect to vocabulary size; recognition in the presence of unknown actions; scenes containing multiple persons; and interactions between multiples persons.

A problem that we could in particular identify for action recognition is the lack of widely-available, realistic datasets. Working with true surveillance footage, sport recordings, movies, and video data from the Internet, can help shift focus to the important open issues mentioned above. This has been well understood by the community, and recently several very realistic datasets have been published, for instance using short clips from Hollywood movies [60] or Youtube videos [66]. Those data sets are challenging the state of the art reported in this survey. Currently, approaches that model spatial and temporal statistics over local feature points are showing the most promising results on those difficult sequences. This is mostly because they do not depend on person or body part detection, which is especially challenging in completely uncontrolled scenes. However, future work needs to come up with more robust feature point detectors and descriptors, and efficient ways to incorporate spatial and temporal structure into the statistics over the detected points (over short term and long term). Some conceptually promising ideas have been already presented [82, 111, 142, 60, 108, 76, 107]. In our taxonomy, such methods fall under the categories of *bags of trajectories*, feature templates and *bags of events*.

We also believe that other representations discussed in this survey will remain important, especially in more controlled environments, such as for instance human-computer interaction and video surveillance, where camera parameters can be controlled and background models can be learned more easily; and entertainment applications, where the positions and appearances of actors are usually known in advance. Body or image models in combination with temporal templates or grammars provide efficient solutions to model temporal and spatial structure of actions in such scenarios. Their success will however strongly depend on much-needed progress in body part detection and tracking.

References

[1] A. Agarwal, B. Triggs, Recovering 3d human pose from monocular images, Transactions on Pattern Analysis and Machine Intelligence 28 (1) (2006) 44–58.

[2] M. Ahmad, S.-W. Lee, Hmm-based human action recognition using multiview image sequences, in: International Conference on Pattern Recognition, vol. 1, 2006, pp. 263–266.

[3] S. Belongie, J. Malik, Matching with shape contexts, in: IEEE Workshop on Content-based Access of Image and Video Libraries, 2000, pp. 20–26.

[4] A. Bissacco, A. Chiuso, Y. Ma, S. Soatto, Recognition of human gaits, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2001, pp. II–52–II–57 vol.2.

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

[6] A. Bobick, Y. Ivanov, Action recognition using probabilistic parsing, in: Conference on Computer Vision and Pattern Recognition, 1998, pp. 196–202.

[7] A. F. Bobick, J. W. Davis, The recognition of human movement using temporal templates, Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.

[8] R. Bodor, B. Jackson, O. Masoud, N. Papanikolopoulos, Image-based reconstruction for view-independent human motion recognition, in: International Conference on Intelligent Robots and Systems, vol. 2, 2003, pp. 1548–1553 vol.2.

[9] O. Boiman, M. Irani, Detecting irregularities in images and in video, in: International Conference on Computer Vision, vol. 1, 2005, pp. 462–469 Vol. 1.

[10] M. Brand, Shadow puppetry, in: International Conference on Computer Vision, 1999, pp. 1237–1244.

[11] M. Brand, V. Kettnaker, Discovery and segmentation of activities in video, Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 844–851.

[12] M. Brand, N. Oliver, A. Pentland, Coupled hidden markov models for complex action recognition, in: Conference on Computer Vision and Pattern Recognition, 1997, pp. 994–999.

[13] C. Bregler, Learning and recognizing human dynamics in video sequences, in: Conference on Computer Vision and Pattern Recognition, 1997, pp. 568–574.

[14] A. Briassouli, V. Tsiminaki, I. Kompatsiaris, Human motion analysis via statistical motion processing and sequential change detection, Journal on Image and Video Processing 2009.

[15] L. W. Campbell, D. A. Becker, A. Azarbayejani, A. F. Bobick, A. Pentland, Invariant features for 3-d gesture recognition, in: International Conference on Automatic Face and Gesture Recognition, 1996, pp. 157–163.

[16] L. W. Campbell, A. F. Bobick, Recognition of human body motion using phase space constraints, in: International Conference on Computer Vision, 1995, pp. 624–630.

[17] S. Carlsson, J. Sullivan, Action recognition by shape matching to key frames, in: Workshop on Models versus Exemplars in Computer Vision, 2001.

[18] I. Cohen, H. Li, Inference of human postures by classification of 3d human body shape, in: IEEE International Workshop on Analysis and Modeling of Faces and Gestures, 2003, pp. 74–81.

[19] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: ECCV Workshop on Statistical Learning in Computer Vision, vol. 1, 2004, p. 22.

[20] R. Cutler, M. Turk, View-based interpretation of real-time optical flow for gesture recognition, in: International Conference on Automatic Face and Gesture Recognition, 1998, pp. 416–421.

[21] F. Cuzzolin, A. Sarti, S. Tubaro, Action modeling with volumetric data, in: International Conference on Image Processing, vol. 2, 2004, pp. 881–884 Vol.2.

[22] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 886–893 vol. 1.

[23] T. Darrell, A. Pentland, Space-time gestures, in: Conference on Computer Vision and Pattern Recognition, 1993, pp. 335–340.

[24] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: International Workshop on Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[25] A. A. Efros, A. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: International Conference on Computer Vision, 2003, pp. 726–733.

[26] A. M. Elgammal, V. D. Shet, Y. Yacoob, L. S. Davis, Learning dynamics for exemplar-based gesture recognition, in: Conference on Computer Vision and Pattern Recognition, 2003, pp. 571–578.

[27] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, X. Twombly, Vision-based hand pose estimation: A review, Computer Vision and Image Understanding 108 (1-2) (2007) 52–73.

[28] A. Farhadi, M. K. Tabrizi, Learning to recognize activities from the wrong view point, in: European Conference on Computer Vision, 2008, pp. 154–166.

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

[30] Z. Feng, T.-J. Cham, Video-based human action classiffcation with ambiguous correspondences, in: Conference on Computer Vision and Pattern Recognition, 2005, p. 82.

[31] R. Filipovych, E. Ribeiro, Learning human motion models from unsegmented videos, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[32] M. Fischler, R. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on Computers 22 (1) (1973) 67–92.

[33] D. Gavrila, L. Davis, Towards 3-d model-based tracking and recognition of human movement, in: International Workshop on Face and Gesture Recognition, 1995, pp. 272–277.

[34] A. Gilbert, J. Illingworth, R. Bowden, Scale invariant action recognition using compound features mined from dense spatio-temporal corners, in: European Conference on Computer Vision, 2008, pp. I: 222–233.

[35] N. H. Goddard, The interpretation of visual motion: recognizing moving light displays, in: Workshop on Visual Motion, 1989, pp. 212 – 220.

[36] R. D. Green, L. Guan, Quantifying and recognizing human movement patterns from monocular video images-part i: a new framework for modeling human motion., Transactions on Circuits and Systems for Video Technology 14 (2) (2004) 179–190.

[37] A. Gritai, Y. Sheikh, M. Shah, On the use of anthropometry in the invariant analysis of human actions, in: International Conference on Pattern Recognition, vol. 2, 2004, pp. 923–926 Vol.2.

[38] R. Gross, J. Shi, The cmu motion of body (mobo) database, Tech. Rep. CMU-RI-TR-01-18, Robotics Institute, Pittsburgh, PA (June 2001).

[39] G. Guerra-Filho, Y. Aloimonos, A language for human action, Computer 40 (5) (2007) 42–51.

[40] Y. Guo, G. Xu, S. Tsuji, Understanding human motion patterns, in: International Conference on Pattern Recognition, vol. 2, 1994, pp. 325–329.

[41] C. Harris, M. Stephens, A combined corner and edge detector, in: Alvey Conference, 1988, pp. 147–152.

[42] D. Hogg, Model-based vision: A program to see a walking person, Image and Vision Computing 1 (1) (1983) 5–20.

[43] M. B. Holte, T. B. Moeslund, P. Fihl, View invariant gesture recognition using the csem swissranger camera, Int. J. Intell. Syst. Technol. Appl. 5 (3/4) (2008) 295–303.

[44] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man and Cybernetics 34 (2004) 334–352.

[45] N. Ikizler, , D. Forsyth, Searching video for complex activities with finite state models, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[46] N. Ikizler, P. Duygulu, Human action recognition using distribution of oriented rectangular patches, in: Workshop on HUMAN MOTION Understanding, Modeling, Capture and Animation, 2007.

[47] N. Ikizler, P. Duygulu, Histogram of oriented rectangles: A new pose descriptor for human action recognition, Image and Vision Computing 27 (10) (2009) 1515–1526.

[48] Y. A. Ivanov, A. F. Bobick, Recognition of visual activities and interactions by stochastic parsing, Transactions on Pattern Analysis and Machine Intelligence 22 (8) (2000) 852–872.

[49] H. Jhuang, T. Serre, L. Wolf, T. Poggio, A biologically inspired system for action, in: International Conference on Computer Vision, 2007.

[50] G. Johansson, Visual perception of biological motion and a model for its analysis, Perception & Psychophysics 1414 (2) (1973) 201–211.

[51] I. Junejo, E. Dexter, I. Laptev, P. Perez, Cross-view action recognition from temporal self-similarities, in: European Conference on Computer Vision, Marseille, France, 2008.

[52] K. Kahol, P. Tripathi, S. Panchanathan, T. Rikakis, Gesture segmentation in complex motion sequences, in: International Conference on Image Processing, vol. 2, 2003, pp. II–105–8 vol.3.

[53] M. Kazhdan, Shape representations and algorithms for 3d model retrieval, Ph.D. thesis, Princeton University (April 2004).

[54] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, in: International Conference on Computer Vision, vol. 1, 2005, pp. 166–173.

[55] Y. Ke, R. Sukthankar, M. Hebert, Event detection in crowded videos, in: International Conference on Computer Vision, 2007.

[56] K. M. Kitani, Y. Sato, A. Sugimoto, An mdl approach to learning activity grammars, in: Proc. of the Korea-Japan Joint Workshop on Pattern Recognition (KJPR 2006), 2006.

[57] A. Kläser, M. Marszałek, C. Schmid, A spatio-temporal descriptor based on 3d-gradients, in: British Machine Vision Conference, 2008.

[58] A. Kojima, T. Tamura, K. Fukunaga, Natural language description of human activities from video images based on concept hierarchy of actions., International Journal of Computer Vision 50 (2) (2002) 171–184.

[59] I. Laptev, T. Lindeberg, Space-time interest points, in: International Conference on Computer Vision, 2003, pp. 432–439 vol.1.

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

[61] I. Laptev, P. Perez, Retrieving actions in movies, in: International Conference on Computer Vision, 2007.

[62] L.-J. Li, L. Fei-Fei, What, where and who? classifying events by scene and object recognition, in: International Conference on Computer Vision, 2007, pp. 1–8.

[63] T. Lindeberg, On automatic selection of temporal scales in time-causal scale-space, in: International Workshop on Algebraic Frames for the Perception-Action Cycle, London, UK, 1997, pp. 94–113.

[64] F. Liu, R. Picard, Finding periodicity in space and time, in: International Conference on Computer Vision, 1998, pp. 376–383.

[65] J. Liu, S. Ali, M. Shah, Recognizing human actions using multiple features, in: Conference on Computer Vision and Pattern Recognition, 2008.

[66] J. Liu, J. Luo, M. Shah, Recognizing realistic actions from videos "in the wild", in: Conference on Computer Vision and Pattern Recognition, 2009.

[67] J. Liu, M. Shah, Learning human actions via information maximization, in: Conference on Computer Vision and Pattern Recognition, 2008.

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

[69] F. Lv, R. Nevatia, Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, in: European Conference on Computer Vision, 2006, pp. 359–372.

[70] F. Lv, R. Nevatia, Single view human action recognition using key pose matching and viterbi path searching, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[71] F. Lv, R. Nevatia, M. Lee, 3d human action recognition using spatio-temporal motion templates, in: ICCV Workshop on Human-Computer Interaction, 2005, p. 120.

[72] D. Marr, H. K. Nishihara, Representation and recognition of the spatial organization of three-dimensional shapes, Philosophical Transactions of the Royal Society of London B 200 (1140) (1978) 269–294.

[73] D. Marr, L. Vaina, Representation and recognition of the movements of shapes, Philosophical Transactions of the Royal Society of London B 214 (1982) 501–524.

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

[75] H. Meng, N. Pears, C. Bailey, A human action recognition system for embedded computer vision application, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[76] R. Messing, C. Pal, H. Kautz, Activity recognition using the velocity histories of tracked keypoints, in: International Conference on Computer Vision, 2009.

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

[78] L.-P. Morency, A. Quattoni, T. Darrell, Latent-dynamic discriminative models for continuous gesture recognition, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[79] P. Morguet, M. Lang, Spotting dynamic hand gestures in video image sequences using hidden markov models, in: International Conference on Image Processing, 1998, pp. 193–197 vol.3.

[80] P. Natarajan, R. Nevatia, View and scale invariant action recognition using multiview shape-flow models, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[81] J. Niebles, H. Wang, H. Wang, L. Fei Fei, Unsupervised learning of human action categories using spatial-temporal words, in: British Machine Vision Conference, 2006, p. III:1249.

[82] J. C. Niebles, L. Fei-Fei, A hierarchical model of shape and appearance for human action classification, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[83] S. Niyogi, E. Adelson, Analyzing and recognizing walking figures in xyt, in: Conference on Computer Vision and Pattern Recognition, 1994, pp. 469–474.

[84] S. Nowozin, G. Bakir, K. Tsuda, Discriminative subsequence mining for action classification, in: International Conference on Computer Vision, 2007.

[85] A. Ogale, A. Karapurkar, G. Guerra-Filho, Y. Aloimonos, View-invariant identification of pose sequences for action recognition., in: VACE, 2004.

[86] V. Parameswaran, R. Chellappa, View invariants for human action recognition, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2003, pp. II–613–19 vol.2.

[87] S. Park, J. K. Aggarwal, Recognition of two-person interactions using a hierarchical bayesian network, in: ACM SIGMM International Workshop on Video Surveillance, 2003, pp. 65–76.

[88] V. I. Pavlovic, R. Sharma, T. S. Huang, Visual interpretation of hand gestures for human-computer interaction: a review, Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 677–695.

[89] P. Peursum, H. Bui, S. Venkatesh, G. West, Human action segmentation via controlled use of missing data in hmms, in: International Conference on Pattern Recognition, vol. 4, 2004, pp. 440–445 Vol.4.

[90] P. Peursum, S. Venkatesh, G. West, Tracking-as-recognition for articulated full-body human motion analysis, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–8.

[91] P. Peursum, G. West, S. Venkatesh, Combining image regions and human activity for indirect object recognition in indoor wide-angle views, in: International Conference on Computer Vision, vol. 1, 2005, pp. 82–89 Vol. 1.

[92] M. Pierobon, M. Marcon, A. Sarti, S. Tubaro, 3-d body posture tracking for human action template matching, in: International Conference on Acoustics, Speech, and Signal Processing, vol. 2, 2006, pp. II–II.

[93] R. Polana, R. Nelson, Recognition of motion from temporal texture, in: Conference on Computer Vision and Pattern Recognition, 1992, pp. 129–134.

[94] R. Polana, R. Nelson, Low level recognition of human motion (or how to get your man without finding his body parts), in: NAM, 1994.

[95] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1990) 267–296.

[96] D. Ramanan, D. A. Forsyth, Automatic annotation of everyday movements, Tech. Rep. UCB/CSD-03-1262, EECS Department, University of California, Berkeley (Jul 2003).

[97] C. Rao, A. Yilmaz, M. Shah, View-invariant representation and recognition of actions, International Journal of Computer Vision 50 (2) (2002) 203–226.

[98] J. Rittscher, A. Blake, Classification of human body motion, in: International Conference on Computer Vision, 1999, pp. 634–639.

[99] N. Robertson, I. Reid, Behaviour understanding in video: A combined method, in: International Conference on Computer Vision, 2005, pp. 808–815.

[100] M. D. Rodriguez, J. Ahmed, M. Shah, Action mach a spatio-temporal maximum average correlation height filter for action recognition, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[101] G. Rogez, J. Guerrero, J. Martinez del Rincon, C. Orrite Urunuela, Viewpoint independent human motion analysis in man-made environments, in: British Machine Vision Conference, 2006, p. II:659.

[102] M.-C. Roh, H.-K. Shin, S.-W. Lee, S.-W. Lee, Volume motion template for view-invariant gesture recognition, in: International Conference on Pattern Recognition, vol. 2, 2006, pp. 1229–1232.

[103] K. Rohr, Towards model-based recognition of human movements in image sequences, Graphical Model and Image Processing 59 (1) (1994) 94–115.

[104] R. Rosales, S. Sclaroff, Inferring body pose without tracking body parts, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2000, pp. 721–727 vol.2.

[105] J. M. Rubin, W. A. Richards, Boundaries of visual motion, Tech. rep., Massachusetts Institute of Technology, Cambridge, MA, USA (1985).

[106] Y. Rui, P. Anandan, Segmenting visual actions based on spatio-temporal motion patterns, in: Conference on Computer Vision and Pattern Recognition, 2000, pp. 1111–1118.

[107] M. S. Ryoo, J. K. Aggarwal, Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in: International Conference on Computer Vision, 2009.

[108] S. Savarese, A. DelPozo, J. C. Niebles, L. Fei-Fei, Spatial-temporal correlatons for unsupervised action classification, in: IEEE Workshop on Motion and video Computing, 2008, pp. 1–8.

[109] K. Schindler, L. van Gool, Action snippets: How many frames does human action recognition require?, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[110] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: A local svm approach, in: International Conference on Pattern Recognition, 2004, pp. 32–36.

[111] P. Scovanner, S. Ali, M. Shah, A 3-dimensional sift descriptor and its application to action recognition, in: ACM International conference on Multimedia, 2007, pp. 357–360.

[112] S. M. Seitz, C. R. Dyer, View-invariant analysis of cyclic motion, International Journal of Computer Vision 25 (3) (1997) 231–251.

[113] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2005, pp. 994–1000 vol. 2.

[114] M. Sheikh, M. Shah, Exploring the space of a human action, in: International Conference on Computer Vision, vol. 1, 2005, pp. 144–149.

[115] Y. Shen, H. Foroosh, View-invariant action recognition using fundamental ratios, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–6.

[116] Q. Shi, L. Wang, L. Cheng, A. Smola, Discriminative human action segmentation and recognition using semi-markov model, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[117] H. Sidenbladh, M. J. Black, D. J. Fleet, Stochastic tracking of 3d human figures using 2d image motion, in: European Conference on Computer Vision, Springer-Verlag, London, UK, 2000, pp. 702–718.

[118] L. Sigal, M. J. Black, Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Tech. rep., Brown University (2006).

[119] C. Sminchisescu, A. Kanaujia, Z. Li, D. Metaxas, Conditional models for contextual human motion recognition, in: International Conference on Computer Vision, vol. 2, 2005, pp. 1808–1815 Vol. 2.

[120] R. Souvenir, J. Babbs, Learning the viewpoint manifold for action recognition, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–7.

[121] T. Starner, A. Pentland, Real-time american sign language recognition from video using hidden markov models, in: International Symposium on Computer Vision, 1995, pp. 265–270.

[122] S. Sumi, Upside-down presentation of the johansson moving light-spot pattern, Perception 13 (3) (1984) 283–286.

[123] T. Syeda-Mahmood, M. Vasilescu, S. Sethi, Recognizing action events from multiple viewpoints, in: EventVideo01, 2001, pp. 64–72.

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

[125] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, International Journal of Computer Vision 9 (2) (1992) 137–154.

[126] D. Tran, A. Sorokin, Human activity recognition with metric learning, in: European Conference on Computer Vision, 2008.

[127] P. Turaga, A. Veeraraghavan, R. Chellappa, Statistical analysis on stiefel and grassmann manifolds with applications in computer vision, in: Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[128] A. Veeraraghavan, R. Chellappa, A. Roy-Chowdhury, The function space of an activity, in: Conference on Computer Vision and Pattern Recognition, vol. 1, 2006, pp. 959–968.

[129] S. Vitaladevuni, V. Kellokumpu, L. Davis, Action recognition using ballistic dynamics., in: Conference on Computer Vision and Pattern Recognition, 2008, p. 8 p.

[130] L. Wang, D. Suter, Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model, in: Conference on Computer Vision and Pattern Recognition, 2007.

[131] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, T. Darrell, Hidden conditional random fields for gesture recognition, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1521–1527.

[132] T.-S. Wang, H.-Y. Shum, Y.-Q. Xu, N.-N. Zheng, Unsupervised analysis of human gestures, in: Pacific Rim Conference on Multimedia, Springer-Verlag, London, UK, 2001, pp. 174–181.

[133] Y. Wang, H. Jiang, M. Drew, Z.-N. Li, G. Mori, Unsupervised discovery of action classes, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 1654–1661.

[134] Y. Wang, P. Sabzmeydani, G. Mori, Semi-latent dirichlet allocation: A hierarchical model for human action recognition, in: Workshop on HUMAN MOTION Understanding, Modeling, Capture and Animation, 2007.

[135] D. Weinland, E. Boyer, Action recognition using exemplar-based embedding, in: Conference on Computer Vision and Pattern Recognition, 2008.

[136] D. Weinland, E. Boyer, R. Ronfard, Action recognition from arbitrary views using 3d exemplars, in: International Conference on Computer Vision, 2007.

[137] D. Weinland, R. Ronfard, E. Boyer, Automatic discovery of action taxonomies from multiple views, in: Conference on Computer Vision and Pattern Recognition, 2006.

[138] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2-3) (2006) 249–257.

[139] A. Wilson, A. Bobick, Learning visual behavior for gesture analysis, in: International Symposium on Computer Vision, 1995, pp. 229–234.

[140] A. D. Wilson, A. F. Bobick, Parametric hidden markov models for gesture recognition, Transactions on Pattern Analysis and Machine Intelligence 21 (9) (1999) 884–900.

[141] S.-F. Wong, R. Cipolla, Extracting spatiotemporal interest points using global information, in: International Conference on Computer Vision, 2007, pp. 1–8.

[142] S.-F. Wong, T.-K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.

[143] Y. Yacoob, M. Black, Parameterized modeling and recognition of activities, in: International Conference on Computer Vision, 1998, pp. 120–127.

[144] J. Yamato, J. Ohya, K. Ishii, Recognizing human action in time-sequential images using hidden markov model, in: Conference on Computer Vision and Pattern Recognition, 1992, pp. 379–385.

[145] P. Yan, S. M. Khan, M. Shah, Learning 4d action feature models for arbitrary view action recognition, in: Conference on Computer Vision and Pattern Recognition, 2008.

[146] M.-H. Yang, N. Ahuja, Recognizing hand gesture using motion trajectories, in: Conference on Computer Vision and Pattern Recognition, vol. 1, 1999, pp. –472 Vol. 1.

[147] A. Yilmaz, M. Shah, Actions sketch: A novel action representation, in: Conference on Computer Vision and Pattern Recognition, 2005, pp. I: 984–989.

[148] A. Yilmaz, M. Shah, Recognizing human actions in videos acquired by uncalibrated moving cameras, in: International Conference on Computer Vision, 2005, pp. 150–157.

[149] L. Zelnik-Manor, M. Irani, Event-based video analysis, in: Conference on Computer Vision and Pattern Recognition, 2001.

[150] Z. Zhang, Y. Hu, S. Chan, L.-T. Chia, Motion context: A new representation for human action recognition, in: European Conference on Computer Vision, 2008, pp. 817–829.

[151] T. Zhao, R. Nevatia, 3d tracking of human locomotion: a tracking as recognition approach, in: International Conference on Pattern Recognition, vol. 1, 2002, pp. 546–551.

[152] W. Zhao, R. Chellappa, P. J. Phillips, A. Rosenfeld, Face recognition: A literature survey, ACM Computing Surveys 35 (4) (2003) 399–458.

[153] H. Zhong, J. Shi, M. Visontai, Detecting unusual activity in video, in: Conference on Computer Vision and Pattern Recognition, vol. 2, 2004, pp. II–819–II–826 Vol.2.