University of Osnabrück
Bachelor Thesis

View-independent human motion analysis of eigensubspace and temporal self-similarity representations

Author: Florian P. Mahner

Primary supervisor: Pattreeya Tanisaro, M.Sc.
Secondary supervisor: Prof. Dr. Gordon Pipa

University of Osnabrück
Faculty of Human Sciences
Department of Cognitive Science

October 26, 2016

Abstract

View-independent human activity recognition has grown into an important research field that has many applications in robotics, video-surveillance or human-machine interactions, among many others. Given an increase in available data it becomes necessary to identify information that preserves innate action structures for representations that are lower-dimensional than the original recordings. The aim of this thesis was to analyze and extend two recognition paradigms for human motions. A focus was appointed to inspecting their view-independent and view-normalized behavior as well as general usability. The first framework models landmark trajectories of skeletal representations using manifold learning techniques. The second paradigm detects motion characteristic changes in the temporal dynamics of action features and describes these as gradient peaks of a self-similarity representation. A degradation of performances was observed with increasing angular differences to closest training samples for the first paradigm. View-normalization reduced spatial variance movements for complex motions and effected greater across-view stability. A consistency of performances across angular differences was discovered for the second framework. A hypersensitivity of parametric analysis was disclosed, with no tendency for automatic global optimization behavior. Normalizing the view did not causally effect view-independent behavior of similarity representations, just its analysis.

Contents

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Consider the following scenario. An elderly woman suffers from Alzheimer's disease at an early stage. She lives on her own, far away from the rest of her family, and takes medications to reduce the cognitive impairment. She wakes up every morning, walks into the kitchen, prepares some breakfast, and uses the stove to make herself a cup of tea. Eventually, an automated voice mildly reminds her to turn off the stove and to take her pharmaceuticals. Later that day, her son logs into the intelligent home system and observes that his mother has taken the medications on schedule and eaten regularly. This assures him that she is still capable of managing daily life on her own.

Automatic human action understanding has become increasingly important regarding numerous everyday fields, such as smart health care, gesture recognition, video surveillance systems, and robotics. Humans have the powerful ability to perceive actions purely from visual information. In doing so, they have a reliable recognition capability that overcomes difficulties such as highly articulated motions, human-object interactions, large inter-class variabilities, and different temporal scales. Over the course of time, action recognition has been divided into a two-class taxonomy (Cedras and Shah, 1994; Poppe, 2010; Wang, 2016). Finding motion- and action-related low-level information for separate sequences has been seen a necessary initialization step for any recognition model. Assigning meaningful semantic information, usually in the sense of action verbs, by matching new inputs against previously trained models, has then been named to be the actual recognition task. The first, so-called feature detection stage is in consequence responsible for finding action characteristic, spatiotemporal information that fit the proposed model best. The following classification stage strives at identifying patterns within these low-level features. Its objective is to obtain a higher-level representation that is similar for semantically related actions and distinct for others.

View-independent human motion analysis is the joint problem of finding a representative model that provides reliable outputs at both, the feature and classification stage, under varying perspectives and transformations. Extending two recognition frameworks in order to investigate their general usability, across view-stabilities and view-normalized behavior for classifying human motions are the goals of this thesis. In the following, differences between models for feature selection will be discussed based on previous literature. Next, an overview of action recognition models will be given and set in context to the paradigms used for this thesis. The concept of model-based view-independence will be described subsequently. Afterwards, data preprocessing will be reviewed and the classifiers described. Then, two frameworks are separately introduced and coherently analyzed regarding their view-independent behaviors. A comparative evaluation will be given at the end.

1.1 Action modeling

Finding the right approach to model human motions from videos has been widely debated over the past years. Early psychophysical research made by Johansson (1973) started examining the functioning of human motion perception.

Figure 1.1. Experiment with Moving Light Displays (Figure taken from Johansson, 1973).

His visual interpretations of biological motions with Moving Light Displays (MLD) have revealed a recognition capability for human motion solely from a set of moving 2d markers (Figure 1.1). These findings have resulted in great discussion about the interpretation of MLD stimuli and have led to multiple reassessments about representation schemes of human poses. Overall, two main classes have evolved over the years, high-level and low-level action models, which are described next.

High-level representations

High-level representations, or so-called top-down models, employ a geometrical framework that tries to account for the biological observations obtained from the MLD. It has been stated that human motions can therefore either be represented as connected body segments,

or 3d volumes (Aggarwal and Cai, 1997). Most approaches try to model human postures through connected segments, so-called kinematic trees or stick-figure models (Akita, 1984; Bharatkumar, Daigle, Pandy, Cai, and Aggarwal, 1994). The connections effect to give a hierarchical structure meaning to the individual body joints. These stick-figures have been extended for volumetric representations. They aim at finding a representation that includes the shape of the human body and models entire parts. Previous research has, for example, modeled human motions through cylindrical representations (Hogg, 1983) or rectangular ones (Ramanan and Forsyth, 2003).

Low-level representations

Low level-representations, or bottom-up approaches, try to extract certain low-level features from an image that are most informative for an action, but not necessarily connected to individual body parts. These low-level features are then assembled to form a higher-level action representation. Creating motion silhouettes in environments where background subtraction techniques could easily be applied, describes one of the earliest approaches. These silhouettes are, for instance, used to create Motion History Images (MHI) and Motion Energy Images (MEI). Both are templates that combine motion and shape information of actions (Bobick and Davis, 2001). The Histogram of Oriented Gradients descriptor (Dalal and Triggs, 2005) has been extended for human action recognition (Thurau and Hlavác, 2008; Ikizler-Cinbis, Cinbis, and Sclaroff, 2009), by representing actions by gradient histograms of individual pose-primitives. Efros, Berg, Mori, and Malik (2003) use optical flow information by extracting a set of spatio-temporal motion features that describe actions over local time-periods. The resulting global templates are then used to classify video frames individually. The bag-of-features approach has recently gained much popularity. It initializes by detecting a set of interest points using the SIFT-algorithm (Lowe, 1999), for example, and builds a so-called codebook for these local descriptors. The corresponding word histogram of reoccurring patches is then used as a feature vector for classification.

Structural representations

Experiments with MLD's have requestioned the structural representation that humans create when perceiving motion. It has been argued that motion information obtained in 2d is either directly used for recognition, or that the stimuli are interpreted after first recovering the 3d motion structure. These controversies have been conveyed onto new recognition frameworks that have been introduced thereafter. Low-level representations have the strong advantage of working independently of any background model. A reconstruction

for limb poses in 3d is not needed, since it can be directly worked with the image data. Due to the high number of degrees of freedom for the human body, in addition to a wide range of poses, it is hardly possible to obtain automatic body-model representations without previous calibration stages that initialize the model. 3d pose estimations of motions have previously been achieved by triangulating multi-camera recordings, or by marker-based human motion capture systems. Whilst the former is sensitive to noise, the latter possesses limited ranges of applications. These findings have shifted earlier research interests towards model-free recognition schemes.

Recent advances in maker-less motion capture systems, like Microsoft Kinect, have nowadays expedited the proceedings towards a kinematic, skeletal representation being more suitable for video-based human action recognition. Skeletal joint configurations are now effortlessly derived and the general practicality of these models therefore increased. Several studies have pointed towards an outperforming of higher-level models against low-level representations (Tran, Kakadiaris, and Shah, 2011; Wang, Liu, Wu, and Yuan, 2012; Wang, Liu, and Wu, 2014). Despite several constraints, like range limitations for depth cameras, or difficulties in consistently track joint locations for occluded or multi-person scenes, actual research indicates promising future results. The presented study therefore analyzes classification effects for 3d human motion capture data.

1.2 Action representation

Once a spatial model has been initialized, motions are characterized as four dimensional objects, depending on the number of features detected at every frame and their corresponding length of recordings. This implies the necessity to detect temporal patterns that have the descriptive power to model the dynamics of action features over time. Several generative approaches have been introduced to model the temporal dynamics of actions, such as Hidden Markov Models (HMM) for 3d points on a motion silhouette (Li, Zhang, and Liu, 2010) or Recurrent Neural networks (Martens and Sutskever, 2011), alongside many others. Some are depending much on the type of feature recognition scheme, whilst others claim to have stable models, independent of the form of found features.

All approaches have to account for the following difficulties: Action durations are mostly not fixed, which results in different frame rates and therefore changing temporal alignments on when an action is performed. Large intra-class varieties among different subjects performing the same action induce severe fluctuations on locally detected features.

Actions are performed at different speeds with diverging local action traits, which can be of importance if their investigated time period is too short. Above all, it can be concluded that any classification scheme has to account for these constraints and provide reliable descriptors for changing temporal, structural and spatial dynamics. What follows introduces the action descriptions selected for this study and related work.

Eigensubspaces

First described in the study *"Analyzing the Subspaces Obtained by Dimensionality Reduction for Human Action Recognition from 3d Data"*, Körner and Denzler (2012) have introduced a novel human motion recognition scheme that utilizes several manifold learning techniques in order to capture the directions of largest variance for temporal action features. They propose a joint recognition scheme, based on the basis vectors themselves obtained from these dimensionality reduction techniques and average action features. This is because finding manifolds of high-dimensional motion recordings has been seen as a necessary step towards a compact and robust feature representation, where action matching becomes less computationally expensive and can be performed in real-time.

Previous literature has intensively investigated the role of eigenspace representations to model the temporal dynamics of low-level and high-level features detected. Huang, Harris, and Nixon (1998) have formerly mapped motion silhouettes obtained from monocular sequences onto a lower-dimensional representation using principal component analysis. A parametric description has been proposed that models the temporal change of image sequences as a change in the corresponding trajectory in the eigenspace (Murase and Sakai, 1996). Comparable work to Körner and Denzler (2012) has been proposed, by projecting action sequences onto their corresponding subspaces for 3d data (Bottino, De Simone, and Laurentini, 2007).

The eigensubspace approach from Körner and Denzler (2012) discards the projection of actions onto their subspace. It takes the projection parameters, or eigenvectors, themselves as form of representation instead. It has been designed to be working in an environment with already tracked joint locations, and claims have only been made for the 3d case. View-independent behavior has been discarded in the analysis of their study, and is described later, in the corresponding chapter of this thesis.

Self-similarity representations

In the study *"View-Independent Action Recognition from Temporal Self-Similarities"*, Junejo, Dexter, Laptev, and Perez (2011) introduce a novel recognition framework, based on similarities that action sequences display over time. They claim to have a recognition framework with high stability across views and model-independent feature analysis. For any given low- or high-level representation of features, they construct a similarity descriptor based on distances of extracted features for all frame-pairs from the recording. The temporal dynamics of action features are then described as histograms of local gradient peaks obtained from the self-similarity matrix.

Several other studies have priorly investigated motion analysis that discards modeling specific image representations or action dynamics. Rao, Yilmaz, and Shah (2002) introduce a learning and model-free descriptor, that captures action characteristic sub-movements. These so-called dynamic instants, or atomic units, are detected by investigating the spatio-temporal curvature of the corresponding 2d trajectories. Resulting maxima are then stable across instances of the same action and views. Temporal similarities of video segments have been used to identify whether two segments are based on the same motion field (Shechtman and Irani, 2005). Instead of comparing motion labels after classifying them, they propose to correlate image space-time intensities of local image segments. They observe that the resulting descriptor is invariant to color, texture and spatial shifts.

Junejo et al. (2011) have built upon these findings, indicating that modeling recurrent dynamics of scenes has high stabilities for subjects within the same class and varying transformations. In contrast to Körner and Denzler (2012), they have explicitly accounted for view-independent properties.

1.3 View-independence

An important step towards a complete semantic understandings of human actions includes a view-independent modeling at both stages of the recognition taxonomy. Choosing between models is therefore also depending on the stability of recognizing features across views. Two main approaches have been adapted to counter-act effects of varying perspectives (Kazhdan, 2004; Weinland, Ronfard, and Boyer, 2011), view-invariance, and view-normalization. View-normalization seeks to shift all actions into a common canonical coordinate frame, such that they obtain a fixed spatial reference point that is stable across

multiple views. In case of the skeletal models used for this thesis, this implies finding a reference marker of the body model, whose global rotation is then set to zero. Translating this marker to the coordinate system center shifts all actions into a common canonical coordinate frame. View-invariance, on the other hand, tries to find a representation that removes view-independent information completely. As such, it is independent of any view-dependent and spatial features and therefore works consistently under large changes. Since view information mostly also displays action characteristics, view-invariance has often been seen as a trade-off between maximum stability across-views and a minimal loss of discriminative functions.

The thesis at hand investigates the general behavior of two recognition frameworks and searches for view-independent properties. Several action models have been introduced and different action representations described. Both recognition schemes have been presented thereafter. By comparing data in which the sample views are shifted into a common coordinate frame, against samples where motion trajectories are not touched, it is additionally aimed at examining the effect of view-normalization.

Chapter 2

Data preprocessing

In consideration of the objective for this study, data from the CMU Motion Capture (MoCap) library was selected. All motions in this dataset were recorded using 12 Vicon infrared MX-40 cameras, that captured the exact 3d positions of the 41 infrared markers taped to the jumpsuit of the actors. Ten actions (*bend, dance, golf swing, jump forward, jump up, march, run, salsa, walk, walk terrain*) were manually selected, each satisfying the requirement of containing at least eight different trials performing the action. Intentionally, the classes were elected to account evenly for inter-class similarities and differences. This allows an evaluative process compared to the difficulty of the task. Semantically close action classes like walk and run or jump up and jump forward are, for example, from an initial understanding supposed to be much more challenging in differentiation, than distant action classes like bend and golf.

First, the correct read-in of the MoCap data was assured. The 41 markers originally recorded in the CMU Graphics Lab are necessary to fully caption very detailed motion categories, such as hand signals or pantomime actions also entailed in the database. Adequate description of such motions requires the availability of exact movement trajectories for many reference points of the corresponding limb. Considering the action classes that were selected for this study, most of these markers are redundant and therefore dispensable. To achieve an optimal trade-off between a maximal reduction of markers and a minimal loss of core action properties, the amount of skeleton markers was reduced to 15 (see Figure 2.1).

Reducing the joints of the original skeleton representation was achieved by taking the mean of the 3d marker position to be merged. The four head markers originally recorded for the skeleton were, for example, concatenated into one central head marker. This results in a first reduction of the skeleton's dimensionality. Every recorded action can therefore be represented by the following matrix

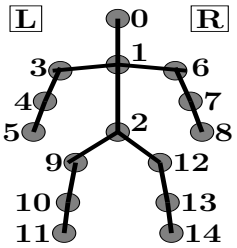


Figure 2.1. Configuration of the skeleton model after the marker reduction. The individual joints are connected to give the action a meaningful visual representation.

$$S = \begin{pmatrix} (x_1^0, y_1^0, z_1^0) & \dots & (x_{N_j}^0, y_{N_j}^0, z_{N_j}^0) \\ \vdots & \vdots & \vdots \\ (x_1^{N_f}, y_1^{N_f}, z_1^{N_f}) & \dots & (x_{N_j}^{N_f}, y_{N_j}^{N_f}, z_{N_j}^{N_f}) \end{pmatrix} \in \mathbb{R}^{(3 \cdot N_j) \times N_f} \tag{2.1}$$

where N_j denotes the number of joints and N_f the number of frames, or time, of the recorded action.

In order to standardize skeleton representations into a coherent spatial framework, all features were rescaled to a range between -1 and 1. Afterwards, data preprocessing was twofold. Skeletons either kept their movement trajectories in space by leaving them as they were read in, or the torso (marker "2" in Figure 2.1) was subtracted from the entire skeleton at each frame. Skeletons of the first set are denoted as moving samples for everything that follows. The second set refers to the actual process of view-normalization, since it translates each action to the coordinate system center. All skeletons then have a rotational invariant reference marker, whose global rotation is set to zero. In the following, these samples are denoted as stationary or rooted samples, since their spatial movement trajectories are removed and all markers kept within the same bounding box. Figure 2.2 depicts an initial skeleton visualization for all actions in the database, after they were view-normalized.

Given that this study aims at investigating rotational behaviors for two human motion frameworks and searches for invariant properties, a coherent rotational space had to be defined, and an axis along which to rotate. Rotating the skeleton along the third dimension (Z) in three-dimensional euclidean space was accomplished by the following matrix multiplication

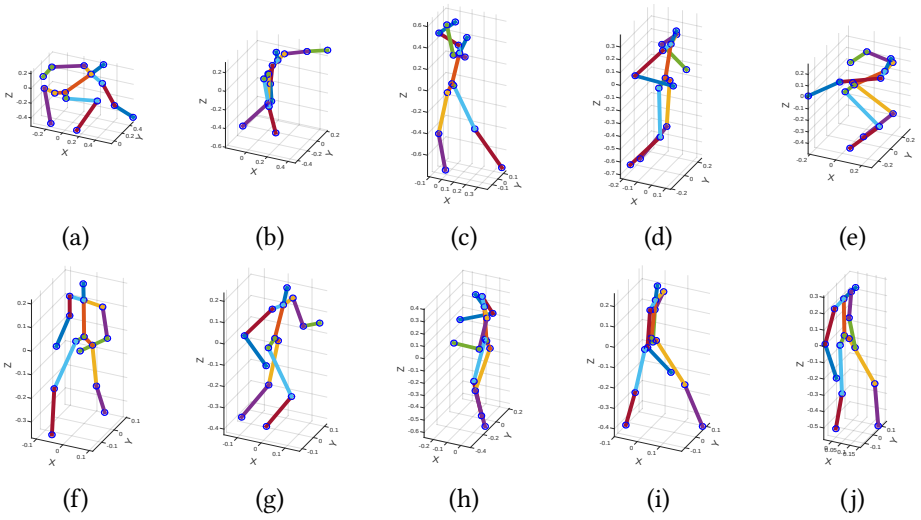
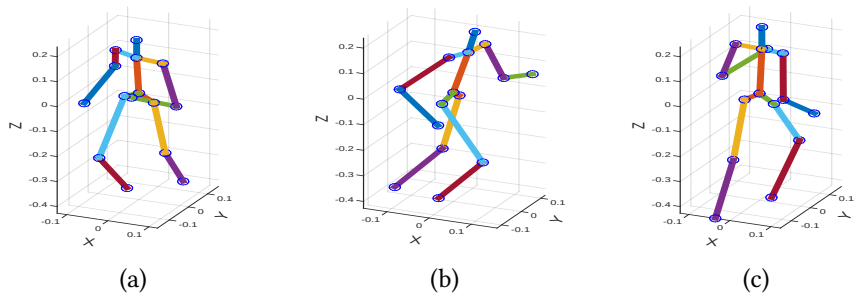


Figure 2.2. Illustration of example skeletons for all action classes. The individual motions are (a) *bend*, (b) *dance*, (c) *golf swing*, (d) *jump forward*, (e) *jump up*, (f) *march*, (g) *run*, (h) *salsa*, (i) *walk*, and (j) *walk terrain*.

$$S_{rot}(\theta) = S_i \times \begin{pmatrix} cos\ \theta & -sin\ \theta & 0 \\ sin\ \theta & cos\ \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \forall i \in N_f \tag{2.2}$$

which rotates an action sequence at every frame i by the angle θ. The rotational reference angle for every action was set to zero degrees, which equals the initial skeleton read-in without applying any rotation in space. As a property of the rotation matrix, a positive angle equals an anti-clockwise and a negative angle a clock-wise rotation along its respective dimension (see Figure 2.3). The rotational scope for classification was limited to a range from -90 to 90 degrees, with 10 degree steps, since a maximal angular range of 180 degrees suffices for most real world applications. Accordingly, 19 different skeleton setups were stored for every action and subject.

All actions were subsequently distributed into a training and testing set for classification. Each subject in the test set had its trained counter-part with equal performer. It was controlled that subjects with identical trials were not reused for classification under a different rotation angle in the test set. Accordingly, the classifiers were fed with a set of training samples, all under the same rotation angles. Depending on the amount of trained angles, test subjects compromised certain minimal angular differences (MAD) to closest training



(a) (b) (c)

Figure 2.3. Rotational scope for a running action ranging from (a) -90 degrees to (b) 0 degrees to (c) 90 degrees.

samples with equal label. The classification results were therefore analyzable in terms of rate developments with regard to increasing minimal angular differences. This allowed the investigation of generalized learning behaviors for individual actions. Depending on the overall performances and the stability or degradation of these results, the individual recognition paradigm was then assessed for view-independent properties.

Chapter 3

Classifiers

This chapter introduces the classifiers that were used in this thesis and kept consistent for the investigation of both recognition frameworks. The k-Nearest Neighbors algorithm and the Support Vector Machine is described in what follows.

3.1 k-Nearest Neighbors

k-Nearest Neighbors (k-NN) is a type of non-parametric, instance-based classifier, that requires a set of labeled training samples. Classifying an unlabeled testing instance initializes by computing its distance to all training samples. A label is subsequently assigned by voting amongst the k closest training samples calculated in the metric space. This study utilized the following distance measure for a training sample x and a testing sample y

$$d_{x,y} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3.1}$$

which is the euclidean distance in n-dimensional space. Unless the Nearest Neighbor algorithm ($k = 1$) is taken, it is advisable to use an odd choice of k to allow voting in all cases. Compared to others, k-NN has the strong advantage of being one of the simplest classifiers that works independently of assumptions about the underlying data distribution.

3.2 Support Vector Machine

The Support Vector Machine (Boser, Guyon, and Vapnik, 1992) (SVM) initializes by mapping its input onto a higher dimensional feature space using a non-linear mapping function. Its core idea is to construct an optimal hyperplane (linear separatix) in this high-dimensional space that is based on samples close to the decision surface (support vectors) and maximizes the margin between individual classes. This hyperplane is then unique and forms the decision boundary for classification (see Figure 3.1).

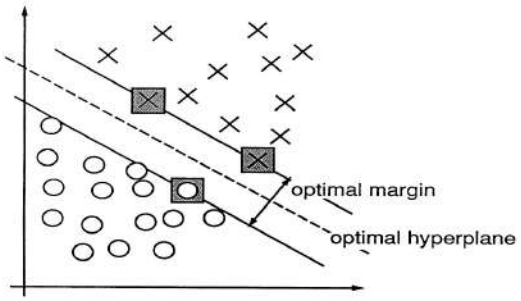


Figure 3.1. Two-dimensional classification problem and the corresponding optimal hyperplane as found by the SVM (Figure taken from Cortes and Vapnik, 1995).

Linear separability

To illustrate this concept, consider a binary classification task $y_i \in \{-1, +1\}$ with a given set $\mathscr{D}$ of n labeled training samples in two-dimensional space

$$\mathscr{D} = \{(x_i, y_i) \mid x_i \in \mathbb{R}^2\}_{i=1}^{n} \tag{3.2}$$

Any hyperplane separating both classes can then be characterized by the following equation

$$\vec{w} \cdot \vec{x} + b = 0 \tag{3.3}$$

where $\vec{w}$ denotes the normal vector perpendicular to the hyperplane. Similarly, the hyperplanes bordering the decision surface are defined as

$$\begin{aligned} \mathcal{H}_1 &: \vec{w} \cdot \vec{x_1} + b = 1 \\ \mathcal{H}_2 &: \vec{w} \cdot \vec{x_2} + b = -1 \end{aligned} \tag{3.4}$$

and the margin is the corresponding space between both these decision functions. Accordingly, it is possible to add the following constraint

$$y_i(\vec{w} \cdot \vec{x_i} + b) \geq 1 \qquad \forall i \in \{1, \ldots, n\} \tag{3.5}$$

for any two hyperplanes to be found, which assures that every data point is situated on the correct side of the maximum margin. Due to the perpendicularity of $\vec{w}$, geometric

transformation reveals that the distance between both margins is $\frac{2}{\|\vec{w}\|}$ since

$$\vec{w} \cdot (x_1 - x_2) = 2 \tag{3.6}$$

$$\frac{\vec{w}}{\|\vec{w}\|} \cdot (x_1 - x_2) = \frac{2}{\|\vec{w}\|} \tag{3.7}$$

Since the goal is to maximize the distance between both margins, this can be reformulated as to minimizing the norm of $\vec{w}$. Solving this minimization problem results in the optimal hyperplane and hence the decision function separating both classes

$$y' = sgn\left(\vec{w} \cdot \vec{x} + b\right) \tag{3.8}$$

Most importantly, this implies that the location of the decision function is fully specified by the small subset of support vectors at the boundaries of the margins. Moreover, this limits classification of a new testing instance to calculating the inner dot product between the two vectors of the decision function.

Classifying the multiple classes of this study was achieved by using the SVM adapting a one-vs-one approach. In this case, the SVM constructs in a first step a set of possible class-pair combinations. This results in $n(n-1)/2$ individual classifiers. At a second stage, it is voted among all class assignments obtained from each classifier and the label with the highest vote assigned. In cases there is a tie, an arbitrary class is usually appointed.

Extension to non-linearity

The main drawback of the Support Vector Machine is its restrictive functioning for linearly separable data. The SVM circumvents this limitation by projecting the original distribution onto a higher dimensional space. The core idea behind this projection is the assumption that any distribution that is not linearly separable in an original m-dimensional feature space can close to always be linearly separated in a n-dimensional feature space, with $n \gg m$. Let $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^n, x \mapsto \phi(x)$ therefore denote the mapping of a point x onto its higher dimensional space. Solving the optimization function by calculating pair-wise dot products for the decision function in high-dimensional space results in a rapid increase of computational complexity. With the help of so-called kernel functions

$$\kappa(x_i, x_j) = \phi(x_i)\phi(x_j)^T \tag{3.9}$$

there exists a possibility to solve the dot product in complexity of the original space, whilst

having a classification accuracy of the transformed higher-order feature space. Achieving to compute inner products in $\mathbb{R}^n$, whilst staying in $\mathbb{R}^m$ is referred to as the kernel trick. This study used the following kernel functions

$\kappa(x, y) = x^T y + c$	(Linear)
$\kappa(x, y) = exp(-\frac{1}{\gamma^2} \parallel x - y \parallel^2)$	(Gaussian radial basis function)
$\kappa(x, y) = (x^T y + c)^d$	(Polynomial)
$\kappa(x, y) = tanh(x^T y)$	(Sigmoid)

Table 3.1. List of kernel functions used for classification with the SVM.

Chapter 4

Eigensubspace representations

The following chapter investigates a novel 3d human motion recognition scheme, as introduced by Körner and Denzler (2012), that models actions based on a combination of average skeleton setups and eigenvectors obtained from various dimensionality reduction techniques. The overall aim of research for this eigensubspace representation was to reassess the validity of the proposed paradigm and to extend its analysis for view-independent behaviors. Moreover, it was sought to make an effective comparison of this scheme between subjects with available motion trajectories and stationary samples and thus to analyze the effect of view-normalization.

The remainder of this chapter is organized as follows. First a detailed description of the constitutive parts used in the paradigm is provided, emphasizing on their combination for a feature vector as action description. Next, classification performances are evaluated for spatial and moving samples and their general view-independent behavior is described. The previously obtained results are then discussed in the context of related literature and an effective conclusion of the general usability is drawn.

4.1 Methodology

4.1.1 Feature extraction

Striving at finding a suitable, time invariant action representation, it was advocated by Körner and Denzler (2012) to use a combined feature vector of the mean shape of an action and its subspaces retrieved from reducing its dimensionality. The mean shape of any skeleton is hereby defined as following

$$S_\mu = \frac{1}{N_f} \sum_{i=1}^{N_f} S_i \tag{4.1}$$

where S_i denotes the skeleton setup at time point i. The mean shape is supposed to characterize and capture the overall movement architecture of an action and provides a suitable representation to discriminate between distant classes. When increasing the difficulty of the classification task, with semantically close actions for instance, the mean shape lacks detailed discrimination properties. It is advanced that capturing the overall variance directions adds the necessary distinctive action properties in this case. These variance directions were obtained by applying several linear and non-linear dimension reduction methods to the skeleton representations, which are described next.

Principal component analysis (PCA)

PCA strives at forming a lower dimensional embedding of an underlying data distribution $\mathscr{D} = \{\vec{x_1}, \vec{x_2}, \ ..., \ \vec{x_i}\}, \ \vec{x_i} \in \mathbb{R}^n$ with zero mean, whilst preserving the maximum of its variance. It thereby finds $m < n$ orthonormal vectors $\{\vec{p_1}, \vec{p_2}, ..., \vec{p_m}\}$, such that the $\vec{p_m}$ are an uncorrelated linear combination of the original distribution and in the direction of largest variance of $\mathscr{D}$. The sample covariance matrix of this distribution is given by $C = \frac{1}{n-1}\mathscr{D}\mathscr{D}^T$. As a property of C, the ith entry on the diagonal represents the variance of the ith variable of $\mathscr{D}$. Given that $C(x_i, x_j) = C(x_j, x_i)$, C is symmetric and hence orthogonally diagonalizable. Therefore, the eigenvalues and eigenvectors of the covariance matrix can be retrieved. The eigendecomposition of the covariance matrix, that is $C\vec{e_m} = \lambda_m \vec{e_m}$, yields $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$ eigenvalues sorted in decreasing order and $\vec{e_1}, \vec{e_2}, ..., \vec{e_m}$ corresponding eigenvectors. These eigenvectors are then called principal components and form a lower-dimensional representation of $\mathscr{D}$.

Kernel principal component analysis (K-PCA)

Since PCA serves as an orthogonal linear transformation, it fails to identify the direction of maximum variances for non-linear manifolds. Kernel PCA (Schölkopf, Smola, and Müller, 1997) provides an extension to non-linearity utilizing the technique of kernel methods, similar to the approach used by the Support Vector Machine. The underlying distribution is similarly mapped onto its higher dimensional space using a kernel function ϕ, where the distribution becomes linearly separable. Instead of extracting the eigenvectors from the covariance matrix $C = \frac{1}{n-1}\phi(\mathscr{D})\phi(\mathscr{D}^T)$ in this new high dimensional space, the kernel trick is applied and the principal components are extracted by solving the eigenproblem of the diagonalized kernel matrix in lower dimensional space. This study made use of the same kernel matrices for K-PCA as previously defined for the SVM (See Table 3.1 on page 16).

Probabilistic PCA (P-PCA)

Probabilistic PCA (Tipping and Bishop, 1999) reformulates the original PCA approach into a maximum-likelihood framework, from which the principal axes emerge. The components are thereby iteratively defined with an expectation maximization (EM) algorithm and a probabilistic density model obtained with corresponding likelihood measure. P-PCA hence has the strong advantage of creating a more efficient version of PCA, that creates an understanding about the underlying data distribution. The combination of a probabilistic model together with EM allows to deal with missing values in the dataset. Moreover, P-PCA can then be used to generatively create new data samples for the distribution using the parameters obtained from the corresponding EM algorithm.

Isometric feature mapping (Isomap)

Whilst PCA aims at finding a data manifold that captures most of the original variance, the Isomap algorithm (Tenenbaum, De Silva, and Langford, 2000), tries to capture the intrinsic geometry of the input space by preserving the interpoint distances for the manifold. When the distance measure is euclidean, this approach is very similar to the one used by PCA. For certain non-linear distributions nevertheless, the euclidean distance cannot provide a suitable representation of data structures. Isomap therefore makes use of an alternative distance measure, the geodesic distance, which estimates the shortest-path distances between points themselves. The algorithm initializes by creating a point graph $\mathscr{G}$. In $\mathscr{G}$ edges are linked together if their euclidean distance falls below a certain threshold (ϵ-Isomap), or if the linked node is one of the k-Nearest Neighbors (k-Isomap). The edges of $\mathscr{G}$ are then weighted with the distance of both nodes. This point graph is subsequently used to compute the shortest path distances between all pairs of points in the graph, using the Floyd algorithm (Floyd, 1962) for example, and stores these in a matrix of proximities $\mathscr{P}$. This matrix is then an estimate of the geodesic distances in the manifold. Transforming $\mathscr{P}$ into a cross-product matrix and solving the eigenproblem like PCA, yields the eigenvectors and eigenvalues of the lower dimensional representation.

Preceding the dimension reduction, the skeleton's dimensionality is determined by the number of individual 3d markers N_m and their number of frames N_f. Reducing the dimensionality yields $\vec{e_m}$ corresponding eigenvectors. The computed lower dimensional representation for all marker characteristics is then of size $N_m \times \vec{e_m}$.

4.1.2 Action representation

The first m basis vectors obtained from one of the dimension reduction techniques were supplemented to the mean shape and formed the action description for this study. Leaving out the projection of the original skeleton onto its subspace, and taking the projection parameters themselves instead, as proposed by Körner and Denzler (2012), results in a vast reduction of computational time required for classifying each sequence at a later stage. Two merging strategies for the eigenvectors were adapted for this thesis and paradigm: Equally adding the basis vectors for the feature vector, removes the qualitative variance difference found within the principal components. All eigenvectors included then have equal weightings inside the feature vector and their different variance contributions are discarded. The eigenvalues found by the dimension reduction method determine the amount of variance covered by their complementary eigenvector. Weighting the basis vectors with their corresponding eigenvalues preserves the variance meaning for the individual components in the feature vector. Higher order principal components thus contribute more to the action description. By comparing these two merging strategies the effect of eigenvalue weighting for stationary and moving samples was additionally examined.

Please note that all eigendecompositions were obtained using the singular value decomposition algorithm. It has been published that this algorithm creates sign ambiguities for the maximal principal component (Bro, Acar, and Kolda, 2008). The possibility of intrinsic sign indeterminacy was removed for this paradigm, by enforcing the maximal principal component to be positive in all cases.

4.2 Results

In order to investigate the behavior of classification performances with increasing angular differences, two different setups were compared. The first setup investigates classification behaviors for moving and stationary samples, when all possible rotation angles entailed in the test set were previously learned on a trained counterpart. This results for all test samples to have a maximal angular difference of 0 degrees. Hence, a first classification comparison between actions with available motion trajectories against stationary processes is conducted, without investigating view-independence at all.

The second setup reduces the amount of training samples for each subject to a maximum of three angles. Empirical investigation revealed an optimal camera setup with best

performances, when the angles were set to -60, 0 and 60 degrees, respectively. This results in four minimal angular differences, since the maximal camera deflection was limited to -90 and 90 degrees. Inevitably, it has to be noted that these differences are unequally distributed for all samples. At 0 degrees, the same number of training angles is only available in the testing set, thus three. For 10 and 20 degrees, six angles are available, and for 30 degrees four.

What follows gives a detailed overview of the performances and methodological parameter settings used in the classification process for the all camera and the three camera setup. The different dimension reduction techniques are thereby coherently compared. The individual classification errors are visualized for moving and stationary samples. A comparison of the different classifiers and explanation of the parameter for the best performing dimension reduction method is given at last.

4.2.1 All camera setup

	Moving		Stationary	
Method	Parameters	Performances (%)	Parameters	Performances (%)
PCA	ω	91.8	–	86
$P-PCA$	ω	87.5	–	87.5
K_{Poly}	$d=3$	**95.3**	$d=3$	**93.5**
K_{RBF}	$\gamma=0.0001$	93.8	$\gamma=0.00001$	93.2
$K_{Sigmoid}$	–	88.7	–	91.6
$Isomap$	$k=5$	90	$k=5$	90.8

Table 4.1. Classification performances and parameter selection summary for moving and stationary samples when all angles were previously trained. The best performing dimension reduction method is highlighted for each dataset in bold.

Table 4.1 depicts and compares the classification performances obtained under the all camera setup for both datasets. The polynomial K-PCA performed best for stationary and moving samples, closely followed by the Gaussian radial basis function (RBF) K-PCA. Whilst PCA and Isomap achieved similar results, Sigmoid K-PCA and P-PCA performed worst for moving samples. Sigmoid K-PCA and Isomap displayed performances similarities and PCA and P-PCA performance difficulties for stationary samples.

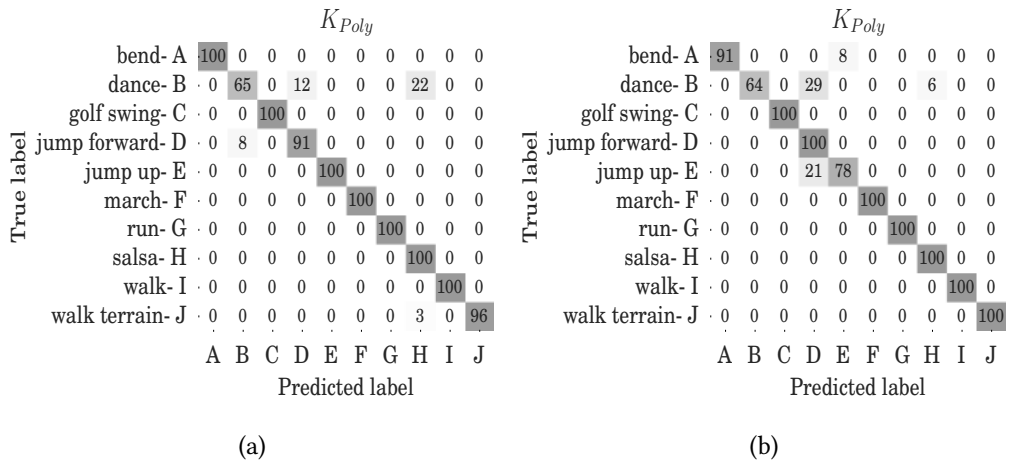


Figure 4.1. Confusion matrix for the classification performances obtained by the polynomial K-PCA for (a) moving samples and (b) stationary samples under the all camera setup.

Given an identical best performing dimension reduction method, the individual classification errors were inspected for the polynomial K-PCA of both datasets (Figure 4.1). It was observed that both datasets had most difficulties classifying *dance* correctly (FNR_{moving} = 37%, $FNR_{stationary}$ = 36%). Besides, minor difficulties were observed for the view-normalized *jump up* (FNR = 21%), solely mistaken for *jump forward*.

4.2.2 Three camera setup

Table 4.2 depicts the classification performances after the amount of training angles was reduced to three cameras. The performance relations between the individual dimension reduction methods are comparable to the all camera setup. Polynomial K-PCA performed best for moving and RBF K-PCA for stationary samples. The overall classification performances of the classifier were subdivided into the individual performances for all possible angular differences.

These angular performances were obtained by dividing the number of misclassification under a certain angular difference by the sum of all samples with that difference. A degradation of performances was observed, correlating with an increase in angular difference. Whilst moving samples indicate little degradation up to the 10 degree benchmark, performances precipitously start dropping from 20 degree onwards. Stationary samples display a comparable behavior, but have greater stability up to 30 degrees.

(a)

Method	Parameters	Performances				
		Overall (%)	0° (%)	10° (%)	20° (%)	30° (%)
PCA	ω	76.5	73.3	75.6	78.3	77.5
$P-PCA$	ω	72.1	74.4	73.3	69.4	72.5
K_{Poly}	$d=3$	**80**	90	91.1	77.2	60
K_{RBF}	$\gamma=0.0005$	76.1	91.1	87.8	71.7	54.2
$K_{Sigmoid}$	–	67.9	91.1	87.2	66.7	46.7
$Isomap$	$k=5$	61.8	76.7	63.9	56.1	55.8

(b)

Method	Parameters	Performances				
		Overall (%)	0° (%)	10° (%)	20° (%)	30° (%)
PCA	–	82.1	80	81.1	83.3	83.3
$P-PCA$	–	78.6	81.1	80.6	77.8	75
K_{Poly}	$d=3$	86.7	93.3	91.7	86.1	75
K_{RBF}	$\gamma=0.00005$	**87**	93.3	91.7	87.2	75
$K_{Sigmoid}$	–	86.7	87.8	87.2	89.4	80.8
$Isomap$	$k=11$	83.5	83.3	85.6	83.3	80.8

Table 4.2. Classification performances and parameter selection summary for (a) moving samples and (b) stationary samples when three angles were previously trained. The best performing dimension reduction method is highlighted for each dataset in bold.

In general, it was found that view-normalization resulted in less performance degradation and therefore outperformed the spatial samples for this setup.

The confusion matrix (Figure 4.2) for the three camera setup undermines what was previously observed. Even though RBF K-PCA performed best in the translated dataset, the minor difference compared to the polynomial K-PCA was for reasons of comparability discarded. The confusion matrix was hence visualized for the polynomial K-PCA in both cases. Similar to the all camera setup, *dance* (FNR_{moving} = 57%, $FNR_{stationary}$ = 36%) showed great difficulties in the classification process. Additionally, major difficulties were observed for stationary samples with *jump up* (FNR = 51%) mostly mistaken for *jump forward* and for moving samples with *golf* (FNR = 30%), *salsa* (FNR = 23%), and *bend* (FNR = 43%).

In summary, it was observed that the recognition scheme shows promising results when view-independence was not investigated at all. Availability of spatial information yields a slightly better action description in this case. Reducing the amount of trained angles

resulted in a degradation of performances with increasing differences and yielded different view-independent behaviors for stationary and moving samples. Overall, it was concluded that stationary samples indicate greater across-view stabilities.

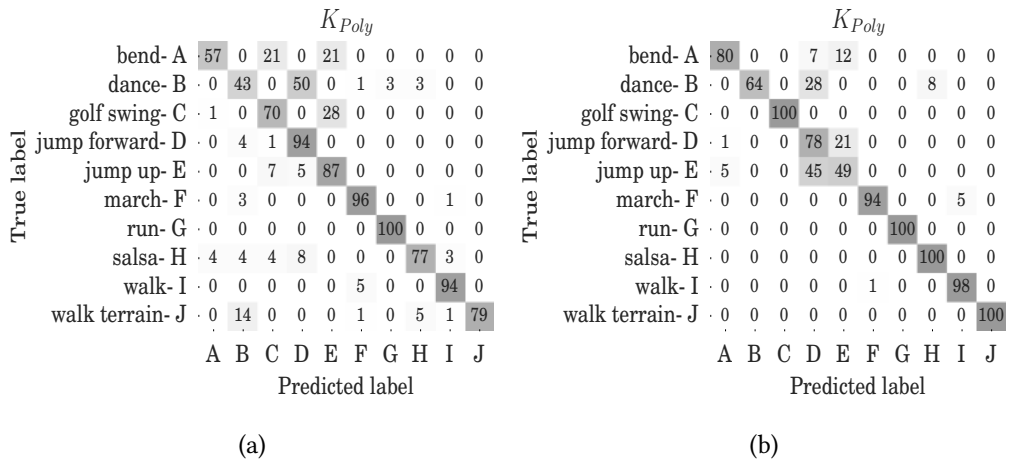


Figure 4.2. Confusion matrix for the classification performances obtained by the polynomial K-PCA for (a) moving samples and (b) stationary samples under the three camera setup.

Please note that a setup with five cameras at angles of -80,-40,0,40 and 80 degrees resulted in similar classification performances up to the 20 degree difference compared to the three camera setup. Polynomial K-PCA again performed best with 91.3% overall performance for the stationary and 89.2% for the moving dataset.

Configurations

Throughout the course of classification, empirical investigation revealed an outscoring of the k-NN classifier over the Support Vector Machine. Varying the number of neighbors whilst keeping all other parameters identical in consecutive trials, exposed highest performances for the nearest neighbor classifier ($k = 1$). Weighting the eigenvectors with their corresponding eigenvalues boosted performances for stationary samples under the all camera setup, whilst for moving samples it was necessary to include all basis vectors equally. Körner and Denzler (2012) investigated the effect of varying the number of principal components in the feature vector. They observed best results when three eigenvectors were included. It was possible to confirm their findings for this framework and the results previously presented were obtained from this setup.

The parameters for the individual methods were chosen to fit the actual camera setup

best. Since the polynomial Kernel-PCA performed best, its degree was analyzed in more detail and contrasted to the one obtained by Körner and Denzler (2012). A third order polynomial performed best in this study. It was assumed that this global minimum is necessary to capture the increased variance distributions resulting from different view-independent variations. A degree of nine, as detected to be best by Körner and Denzler, seems from an initial understanding to overfit the underlying distribution and suits well for single-view samples used in their study. It is nevertheless expected that this degree fails to generalize for most across-view classification tasks.

4.3 Discussion

The aim of the chapter at hand is to investigate the view-independent stability as well as the importance of view-normalization for human motions, represented as a combination of their average skeleton setup and lower-dimensional eigensubspaces. Performances were coherently compared between subjects with available motion trajectories and stationary subjects under two different camera setups. In the end, results indicated comparable performances for both datasets and single views and a degradation of performances correlating with an increase in angular difference. Stationary, or view-normalized, samples exhibited greater across view-stabilities than spatial ones.

What follows targets to explore the general view-independent usability of the proposed scheme for both datasets, and thereby inspects how the individual parts of the feature vector diverge with increasing rotational differences. Furthermore, it is analyzed to what extent across-view stabilities deviate for action classes themselves and how these deviations specifically caused misclassifications for other actions. In the end, it is striven to show that subjects with available motion trajectories display greater inter- and intra-class feature vector fluctuations and that the relative pose-variations of stationary samples are more stable descriptors.

4.3.1 Moving samples

Mean shape

The general spatial locations and corresponding decision boundaries obtained from the k-NN classifier are depicted for the mean shape and all action classes in Figure 4.3. It was recognized that semantically related action classes convey their relatedness onto spatial closeness, as can be observed with *march, run, walk*, and *walk terrain* for instance. A sim-

ilar behavior was noted for *jump up* and *jump forward*. Moreover, it was observed that depending on the type of motion, subjects exhibit different spatial behaviors. Actions were for reasons of analysis subdivided into two motion category types: complex and simple motions. Actions assigned to the complex motion category (*dance, golf swing, salsa, walk terrain*) transmit their complexity either onto larger spatial expansions, or form different clusters across multiple subjects (Figure 4.3 (a)). Simpler motions on the other hand (*bend, jump forward, jump up, march, run, walk*), exhibit greater spatial constancy. The explanation for this is grounded within the definition of the motion itself. Moeslund, Hilton, and Krüger (2006) have previously introduced a motion hierarchy consisting of action primitives, actions and activities. Action primitives refer to atomic entities that describe an action, whilst actions are ordered sequences of action primitives. Activities are then high-level combinations of individual actions. Complex motion categories mostly resemble activities. *Dance*, for example, is constituted of ordered sequences of much simpler sub-motions, like multiple marching and bending actions. Since the variability of continuous alignments of these sub-motions is large, it is explicable that the mean shape of complex actions has less intra-class stability than the one of simpler motions.

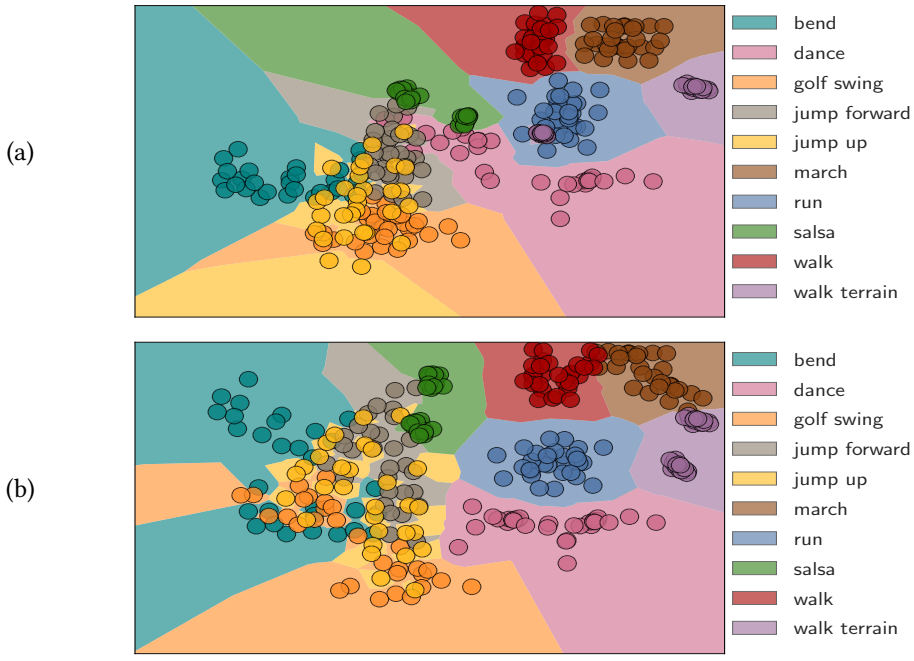


Figure 4.3. Decision boundary of the k-NN classifier and the mean shape for (a) two samples at 0 degrees and (b) one sample at 0 and 30 degrees. The decision surface demonstrates greater spatial movements across angles and subjects for complex motions.

Inspecting the spatial evolution of the mean shape for one subject under different rotation angles revealed a similar behavior compared to multiple subjects under identical rotation angles (Figure 4.3 (b). This extension is due to an increase of possible feature range for complex motions and therefore an increase in spatial displacements with rotations. The decision surface for an action, as created by the classifier, is bounded by the maximum deflection of any possible feature. To illustrate these rotational differences between two action categories, consider the actions *walk* and *dance*. All feature extremities for *walk*, like the position of hands or legs, are in close proximity to the rest of the skeleton body. By contrast, the feature locations of extremities for *dance* are far away from the skeleton and spread along a much greater extension in space, which affects the rotational displacements.

Above all, it can be concluded that the mean shape provides a solid initial action distinction for spatial samples, with a lack of inter-class and view-independent stability, depending on the motion category.

Eigensubspaces

In light of the findings for average skeleton representations it had to be inspected how the basis vectors behave across-views, and if the different categorical spatial movements observed there manifest themselves in similar variance extensions. An exemplary visualization of variance extensions for the first principal component and all actions is depicted in Figure 4.4.

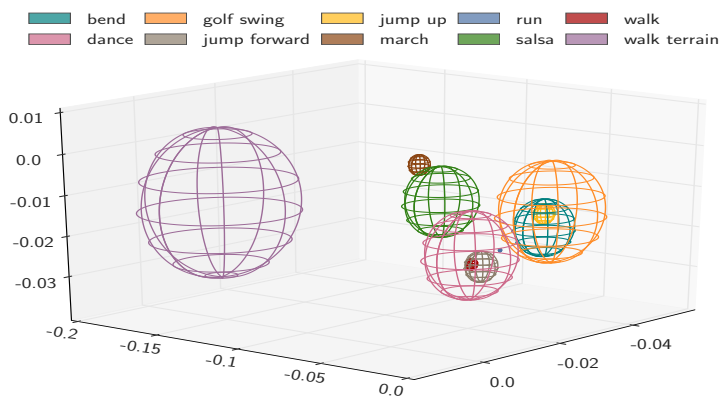


Figure 4.4. Spherical distribution for the variance directions of the first principal component. The two samples considered at 0 degrees indicate greater spatial variance extensions for complex motions compared to simpler motions.

Two random samples for each action at 0 degrees were considered for Figure 4.4, and the sphere plotted around the mean of these samples with a radius equal to the maximal deflection of any feature from that mean. It can be seen that the action specific movement extension observed for the mean shape is preserved for the variance relationships in the linear manifold. The size of the spheres reflect to some degree the complexity of the motion class. The more intra-class variety a motion has, the larger the sphere. Previous research has argued for a spatial distinctiveness of class clusters in their corresponding eigenspace representation (Bottino, De Simone, and Laurentini, 2007). Their visualization of principal components exhibits a very similar behavior compared to the spherical distribution, even though not explicitly mentioned in their study.

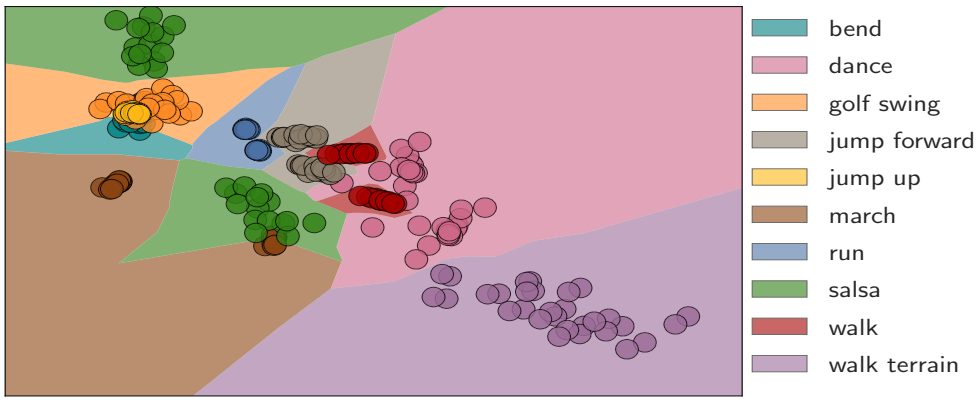


Figure 4.5. Decision boundary for the k-NN classifier of the first principal component of one sample at 0 and 30 degrees. Actions that were misclassified are either located on similar spatial locations (*bend, golf, jump up*), or have a greater spatial extension due to the complexity of the movement class.

The decision boundary, as highlighted in Figure 4.5, shows that the basis vector extension for different samples observed in the spherical distribution transfers to a rotational instability in space. *Jump up, bend*, and *golf*, which were repeatedly misclassified in the three camera setup, are clustered on the same location, which results in a difficult differentiability for the classifier. *Dance, salsa*, and *walk terrain* display, due to their large variability, greater rotational movements also for the first principal component. It can be observed that *dance* is closely located to *jump forward*. Due to the motion complexity of *dance* it is likely that different samples are located within the decision boundary of *jump forward*. Despite having no semantic relation, this reasons why both actions were misclassified for each other with increasing angular differences. The SVM classifier would in this case have a greater

difficulty of learning a conceptual representation of *dance* for a specific angle. Instead, it learns individual samples labeled *"dance"*, which do not necessarily possess a certain spatial relatedness. The k-NN classifier on the other hand has greater probabilities of finding the nearest neighbor in a different action. These behaviors were observed to be comparable across all other angles.

In summary, it can be concluded that subject trajectories themselves, without view-independence, are reasonable predictors for the actions classes investigated in this study. Visualizing the decision boundary for the mean shape and the spherical distribution for the first principal component indicates less stability for complex motions across multiple subjects. A similar behavior was observed across views, announcing a large volatility of the proposed scheme from a rotational difference of 20 degree onwards. Combining the inter-class and rotational instability, especially for complex motions, explains the classification performances of 60% at a 30 degree difference.

Please note that all samples selected for the decision boundary visualizations could have been arbitrarily replaced, without affecting the general deductions made for the analysis. For the sake of visibility, visualization was nevertheless restricted to two samples or angles.

4.3.2 Stationary samples

Translating skeletons in space by subtracting the torso at each frame, shifts the skeleton from a spatial representation into a new coordinate system, where its time-dependent movement information is removed. Instead, each action becomes characterized by relative pose-variations between individual frames.

This results in two main implications for the feature vector components. On the one hand, shifting actions into a unified coordinate system removes the spatial distinctiveness of average skeleton joints. The mean shape then represents relational joint locations inside this spatial framework, which consequences all features to be on similar locations with comparable extensions. On the other hand, eliminating time-related movement direction when normalizing the view, changes the relations of intra-class basis vector scales. The latter is highlighted in Figure 4.6. Depending on whether the eigenvectors are calculated for available spatial trajectories or stationary samples, structural relations of principal components between classes diverge. It was observed that for spatial samples and simple, one-directional, motion categories such as *walk*, for example, the first eigenvector mostly

exhibits time-related movement directions. Notwithstanding that these directions are the essence of the action's variance, it distorts the comparative meaning of higher order eigenvectors for low and high variance actions. Oppositely, for stationary samples, higher order eigenvectors display action characteristic variances, which are time-independent. In Figure 4.6 (a) it can be observed that the first principal components are at a similar scale for all actions, which indicates that the motion complexity of *dance* is not transmitted for the first eigenvector compared to *walk* and *run*. In Figure 4.6 (b), it can be discovered that the categorical motion relationships are conveyed from the first principal component onwards. *Run*, for example, has a small relative variance for the first principal component since the maximum variability of features from that class itself is little, irrespective of time. Oppositely, *dance* has high movements between poses, which results xin high fluctuations of the first component.

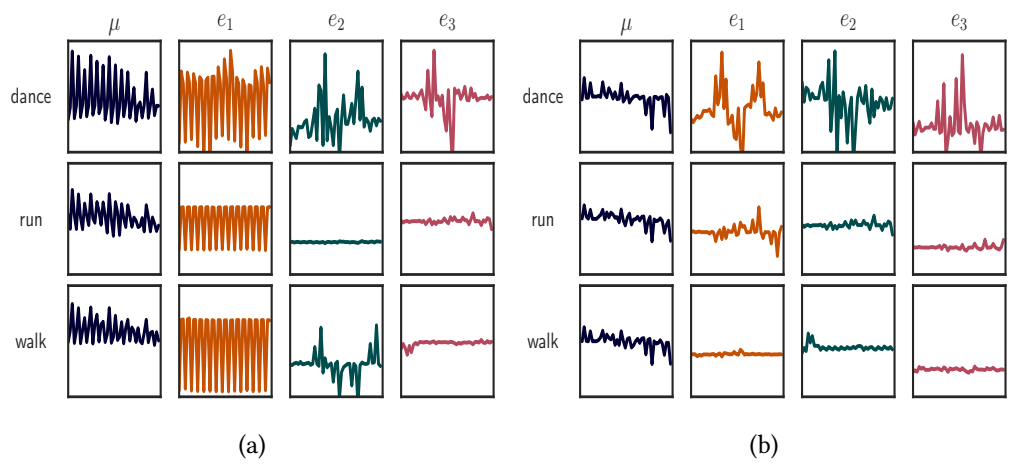


Figure 4.6. Component wise feature vector visualization of *dance*, *run*, and *walk* for (a) spatial samples and (b) translated samples at 0 degrees. All motions were averaged over three samples. Motion complexity is transferred onto higher order eigenvectors for translated samples. The availability of time-related movement direction for spatial samples removes this structural difference for the first principal component.

These findings are in line with the previously obtained results. Recall that classification performances for translated samples increased when the eigenvectors were weighted with their corresponding eigenvalues, whilst spatial samples performed best when the eigenvectors were equally included. In some sense, this represents an enhancement of discovered variance relationships for stationary samples and a necessity for all variances with spatial samples.

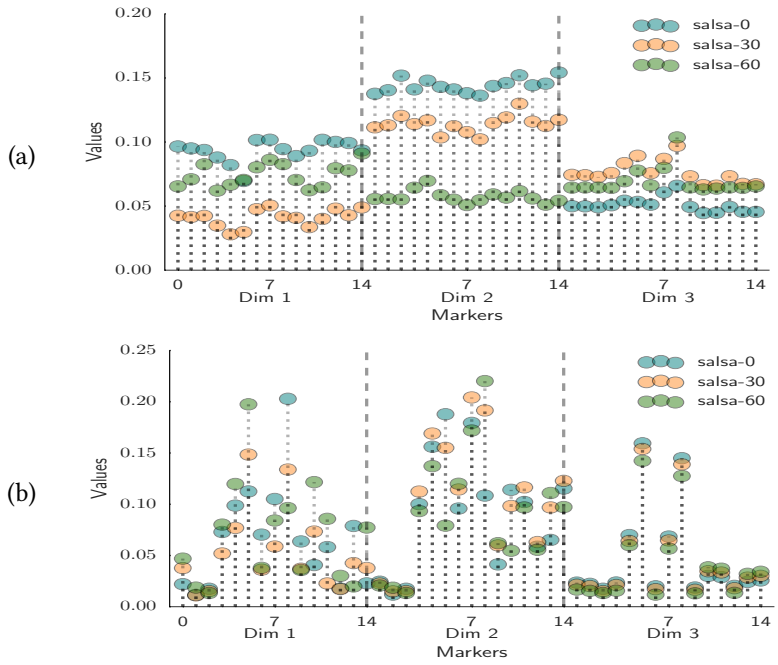


Figure 4.7. Individual value distribution of markers of the first principal component for (a) moving samples and (b) stationary samples and *salsa*. The respective markers were obtained by averaging over all eight subjects. It can be observed that view-normalized samples exhibit less value fluctuation across angles.

Due to the removal of spatial information for the mean shape and different scaling for the subspaces, decision boundaries were not visualizable for stationary samples. Whilst this posits some drawback in terms of a coherent spatial visualization of different view-independent stabilities for moving and stationary samples, the analysis was shifted to a comparison, where the individual value and not spatial extensions are considered for increasing angular differences. Figure 4.7 depicts such a visualization for *salsa*, illustrating the individual marker locations of the first principal component for both moving and stationary samples. It can be observed that fluctuations between marker values are greater for moving and less for view-normalized samples. Whilst in the one case, different spatial locations seem to have large effects on the variance directions discovered, pose-variations exhibit greater stability across angles.

These findings allowed derivations for the general usability of the proposed recognition scheme. In terms of general class distinctiveness moving and stationary samples behaved in

a comparable manner. Eigensubspace representations effected reliable classifications with performances up to 95%. These findings are in line with previous research modeling actions through eigensequences (Bottino, De Simone, and Laurentini, 2007; Wang, Liu, Wu, and Yuan, 2012). The necessity of view-normalization was nevertheless presupposed in all cases, without inspecting its effect for their proposed paradigm. Since moving subjects performed slightly better for the all camera setup, and given that stationary samples already indicated classification difficulties for *jump up* and *jump forward*, it was inferred that spatial information can even be a better predictor in some cases. Yet its success is depending on the actual locations of recordings, which may ease classification in one case, but mostly evokes a bias in terms of general action representations and finding of core descriptive features.

Inspecting the view-independent behavior of the paradigm, it was observed that translated or center-shifted samples are more stable predictors, due to the possibility to obtain more generalized action representations across angles. View-normalization makes the available data independent of the performers characteristics and motion orientation. This global rotation of zero results in a greater across view-stability for eigensequences and therefore more stable action descriptors. Oppositely, it was disclosed that performances dropped for stationary samples at 30 degrees difference to 75%. This indicates that the overall view-independent usability of the paradigm is only valid up to a certain degree of angular difference and raised questions about possible causes of instability. Given that view-normalization removes the spatial bias, action specific variance distributions are enhanced. Despite then having greater across view constancy, it evokes differentiation difficulties for actions with similar variance directions. It remains, nevertheless, open at which stage these difficulties arise. *Walk* and *run*, for example, displayed almost no complications, whilst *jump up* was regularly mistaken for *jump forward* under both camera setups. Having said that, it can be concluded that variance directions are action descriptors that suffer in differentiation for view-normalized samples, if the actions are closely related on a semantic level. This is because this relatedness mostly implies a corresponding variance similarity. On the other hand, it was deduced that actions with available movement trajectories lose their view-independent stability due to a larger spatial uncertainty for complex motions. This behavior manifests itself for the mean shape and the variance directions.

Classifiers

Empirical investigation revealed a coherent outscoring of the nearest neighbor classifier over the SVM. Since both approaches represent two different types of learning, these re-

sults had to be evaluated cautiously. Nearest neighbor primarily performed best, given the idealized motion capture recordings. The 3d depth sensors that were attached to the subjects and from which the skeleton joints were obtained, had close to identical positionings for all actions. If marker locations would not have been fixed but estimated from images through interest point detection algorithms, for instance, this data cleanliness would partially be removed. To counteract the resulting effects of increasing noise and outliers, a vote among an increased number of neighbors would most probably become necessary. k-NN is very much dependent on the actual size of the dataset, and has after some algorithmic improvement, a definite time complexity of $O(\mathscr{D}\mathscr{N}_f)$, where $\mathscr{D}$ is the number of samples in the dataset, and $\mathscr{N}_f$ the dimensionality of each sample. Since it is depending on $\mathscr{D}$, it becomes computationally very expensive for large datasets.

The SVM computes its decision function using a small subset of support vectors that span the hyperplane. It outperforms k-NN for large datasets in terms of memory efficiency. Despite k-NN being greatly influenced by the amount of available samples for every class, it does not include any learning effects. The learning of real class-boundaries that the SVM includes is depending on the amount of training data available. Increasing the number of samples increases the computational complexity. This is due to an increment of required support vectors that are needed to account for the underlying structure. It nonetheless also increases the predictive power of the classifier. It therefore needs to be inspected how both classifiers would behave for larger datasets. For the limited case of the dataset used for this thesis, it can be concluded that k-NN outperformed the SVM.

Chapter 5

Self-similarity representations

The following chapter introduces a model-free human action recognition scheme, as first described by Junejo et al. (2011), which characterizes motions as similarities and dissimilarities of action features over time. The corresponding patterns in the temporal dynamics are analyzed using a classic object detection algorithm. The objective of research for self-similarity representations used in this thesis was to re-evaluate their across-view stability and methodology of analysis, and to explore the effect of view-normalization.

This chapter is organized as follows. First, the concept of modeling recurrent dynamics through self-similarities is explained and the Histogram of Oriented Gradients descriptor introduced. The combination of features for action descriptors is emphasized hereafter. Next, classification performances are provided for moving and stationary samples, under two different feature representations. These results are then evaluated in terms of view-independence and view-normalization, and discussed in light of previous research.

5.1 Methodology

5.1.1 Self-similarity matrix

The idea of a self-similarity matrix (SSM), as introduced by Junejo et al. (2011), was originally derived from the concept of a recurrence plot, which provides a possibility to model and visualize the recurrent dynamics of a system in phase space. In that regard, SSM's are a representation based on similarities of action sequence pairs and defined as following

$$d_{i,j\,=\,[1,\,\ldots\,,T]} = \begin{pmatrix} d_{1,1} & \ldots & d_{1,T} \\ \vdots & \vdots & \vdots \\ d_{T,1} & \ldots & d_{T,T} \end{pmatrix} \quad \in \mathbb{R}^{(T\times T)} \quad \text{with} \quad d_{i,j} = \|d_i - d_j\|_2^2 \tag{5.1}$$

where $d_{i,j}$ denotes the distance between frame i and frame j. Every action is then represented as a square matrix depending on the length of the recording. The euclidean distance measure was chosen between individual frames, where $\|\cdot\|$ denotes the euclidean norm. As a property of the SSM, values on the main diagonal are the distance of a frame to itself and hence zero. This consequences a symmetry along both sides of the diagonal. Each entry in the SSM represents the absolute correlation between frame-pairs of an action description. Self-similarity matrices therefore provide a possibility to model the temporal dynamics of movement trajectories.

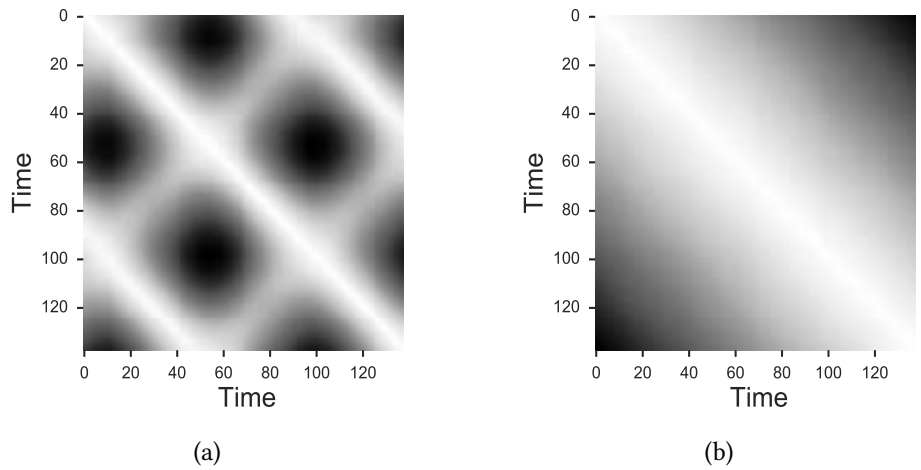


Figure 5.1. Visualization of the self-similarity matrix for one *"running"* action that was (a) view-normalized, or (b) kept as read-in. Please note the structural differences induced by the translation.

A first visualization of these structures is given with Figure 5.1. The dynamic instants of an action are represented as patterns inside the SSM. These patterns represent action specific behaviors within certain temporal expansions. Stationarity of movements is, for example, described by homogeneous areas since the relative change between frames is little (Marwan, Romano, Thiel, and Kurths, 2007). Capturing these patterns can therefore be considered to be a problem of object detection. The Histogram of Oriented Gradients (HOG) descriptor (Dalal and Triggs, 2005) provides a possibility to describe global object appearances, by finding intensity gradient distributions inside local patches. HOG has the strong advantage to avoid reliance on absolute values inside the SSM, whilst still capturing its dynamic patterns. The functioning of the HOG descriptor is described subsequent.

5.1.2 Histogram of Oriented Gradients

Originally introduced for scale-invariant pedestrian detection, HOG operates by forming a global representation of an image through concatenation of local histograms. It does so by transforming any image I into a gradient image ∇I in a two-dimensional convolution operation using $\vec{g_x}$, $\vec{g_y}$

$$\vec{g_x} = \begin{pmatrix} -1 & 0 & 1 \end{pmatrix} \qquad \vec{g_y} = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} \tag{5.2}$$

and thereby identifies the gradient, or relative change of pixels, in the x- and y-direction. The size of the gradient vector determines how many surrounding pixels to consider. The gradient direction is obtained by taking the inverse tangent of both these directions

$$\theta = tan^{-1} \left[\frac{g_x}{g_y} \right] \tag{5.3}$$

The gradient angle θ then point into the direction of the largest rate of intensity increase for the given image point. HOG initializes by dividing the image into smaller cells and calculates for every pixel in that cell its corresponding gradient direction. Resulting angles are then included into an unsigned histogram. The histogram consisted, in the case of this thesis, of nine bins and ranged from 0 to 180 degrees. To counteract possible difficulties of hard assignments for angles close to the individual bin boundaries, the gradient direction is added with equal shares depending on the distance to the closest bin. The output of the HOG descriptor is then delineated by nine binning descriptions for each subsampled point of analysis in the SSM. The locations of analysis where HOG was applied to inside the SSM are described next.

5.1.3 Action representation

Due to a general avoidance of time-series classifications for this study, binning assignments had to be restricted to a consistent number of iterations for all actions. Following Junejo et al. (2011), a feature descriptor was built by applying HOG for equally distributed windows along the diagonal of the SSM. Due to the symmetry of distances in the SSM, shifting the rectangular windows centered on the main diagonal would result in half of the gradient directions being redundant. The windows were therefore right-shifted away from the diagonal, such that the redundant areas are moved outside of the detection area of the descriptor.

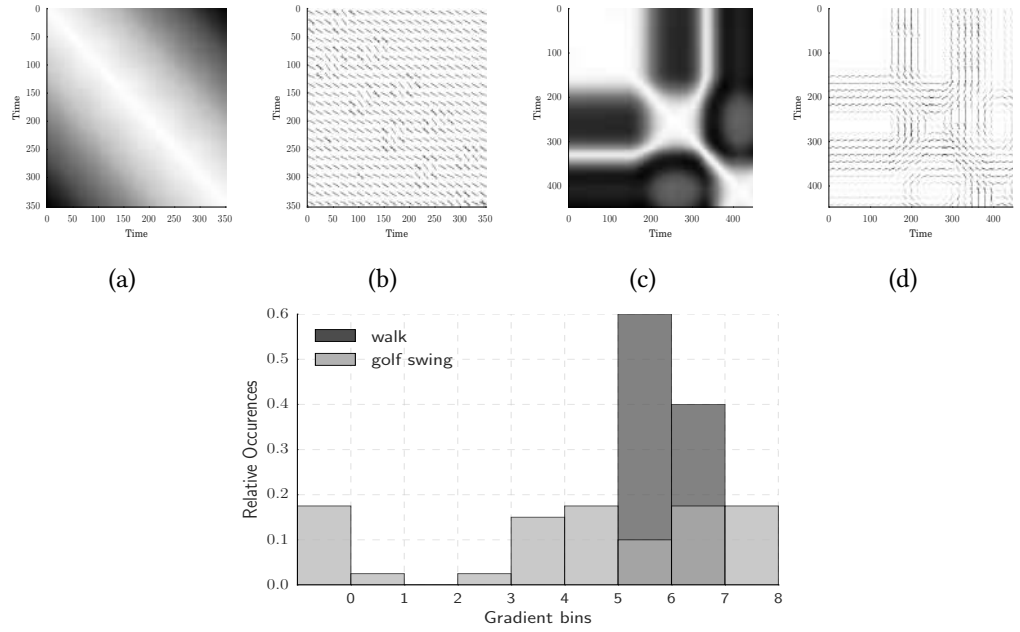


Figure 5.2. Visualization of self-similarity matrix and corresponding gradient orientation as obtained by HOG descriptor. (a) and (b) display the SSM and gradient orientations for *walk*, (c) and (d) display the SSM and gradient orientations for *golf* and (e) the histogram of maximum bin assignments for each point of analysis of both SSM's. It can be observed that the global gradient structure is conveyed onto the nine bins.

At each iteration, the window was divided into four smaller cells. In each of these cells, the nine dominant gradient orientations were computed and stored. Classification was subsequently compared between a feature descriptor containing all iterative binning assignments against an average binning description. The averaged description was obtained by calculating all gradient directions in a first step and afterwards taking the mean gradient binning over all iterations. In the end, this can be seen as a global representation of an action entirely based on local binning structures. This results in a large reduction of computational complexity, since the feature vector length is reduced to nine. Figure 5.2 depicts the self-similarity matrix for *walk* and *golf*, as well as their corresponding gradient images obtained after applying HOG. An application on the entire image revealed gradient orientations similar to the general shape of the SSM. Capturing the global gradient structures that can be observed there, by analyzing local patches, is the main task for this action recognition approach.

5.2 Results

Similar to the analysis of the eigensubspace framework, it was aimed at finding an optimal trade-off between the amount of required training angles and a stability of classification performances. The camera setup was finally reduced to a single training angle at 0 degrees. This resulted in a maximal angular difference of 90 degrees. In the end, performances were independent of the amount of previously trained angular representations and therefore consistent across all minimal angular differences.

This section is organized as follows. Classification results are given for feature descriptors of moving and stationary samples under the all and average binning assignment first. Misclassifications are systematically evaluated for each of the given results. The configuration setups are investigated next, with an emphasis on the window size of the HOG descriptor at each iteration and the number of analysis points considered along the diagonal.

5.2.1 Performances

All binning description

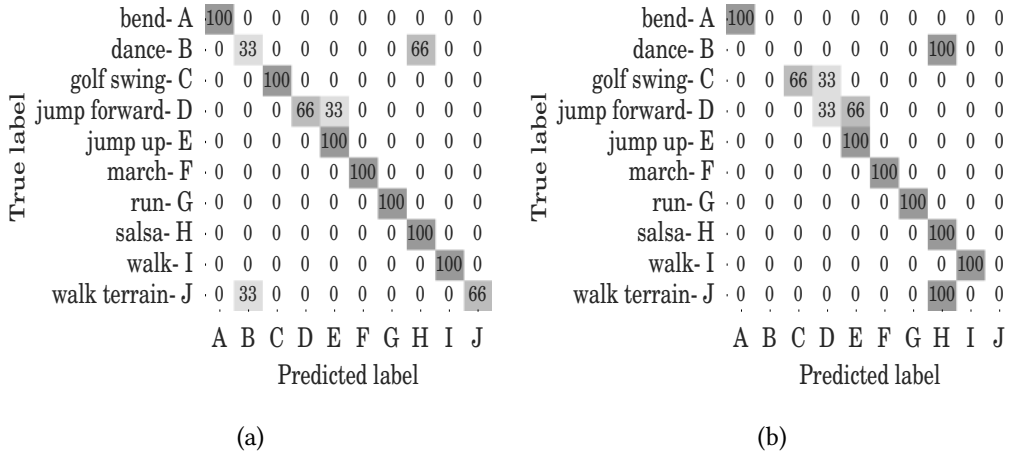


Figure 5.3. Confusion matrix obtained for the all binning classification process for (a) moving samples and (b) stationary samples.

Running classification for feature descriptors entailing all captured gradient orientations revealed peak performances of 86.7% for spatial and 70% for stationary samples, under different configuration setups. Figure 5.3 depicts the individual classification errors for both

datasets. Most difficulties in the classification process for both datasets were observed for *dance* (FNR_{moving} = 66%, $FNR_{stationary}$ = 100%), *walk terrain* (FNR_{moving} = 33%, $FNR_{stationary}$ = 100%), and *jump forward* (FNR_{moving} = 33%, $FNR_{stationary}$ = 66%). It was noted that misclassifications occurred in blocks of 33%. Since the testing set entailed three subjects, this resulted in the first deduction that these misclassifications probably occurred subject wise, irrespective of their angle of rotation.

Average binning description

Running classification for the averaged binning description resulted in performances of 90% for moving and 86.7% for stationary samples, consistent for all angular differences. Figure 5.4 depicts the confusion matrix for both datasets. Misclassification similarly occurred in blocks of 33%. *Dance* (FNR = 33%), *golf* (FNR = 33%), and *jump forward* (FNR = 33%) had most difficulties for moving samples. *Jump forward* (FNR = 66%) and *jump up* (FNR = 66%) were repeatedly mistaken for stationary samples. Opposite to the all binning description, *walk terrain* was classified perfectly.

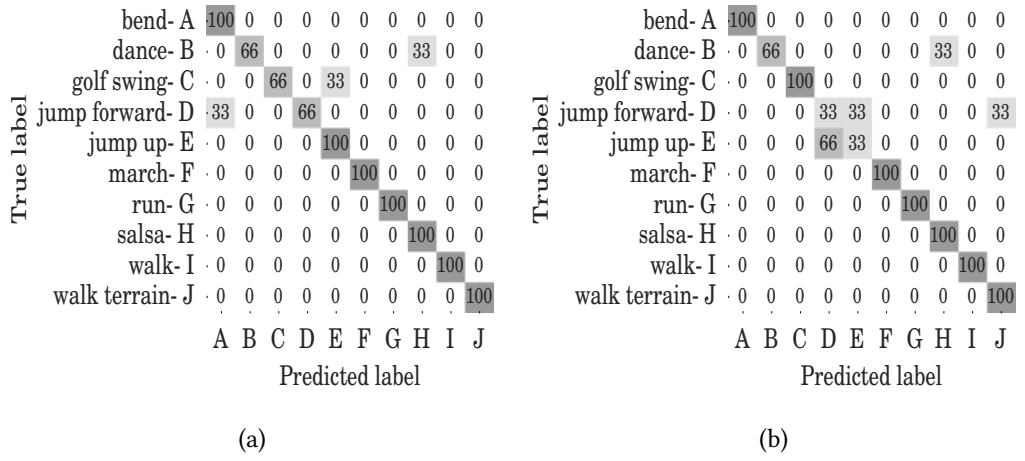


Figure 5.4. Confusion matrix obtained for the average binning classification process for (a) moving samples and (b) stationary samples.

In summary, it was noted that the best performing configuration for moving and stationary samples under the averaged description outperformed the all binning setup. The individual parameter settings of the HOG descriptor that led to the results are described next.

5.2.2 Configurations

Window sizes

Throughout the course of analysis, parameters of the HOG descriptor displayed great sensitivity, which effected large performance fluctuations depending on their selection. It was observed that performances were deeply correlating with actual window sizes considered, and less dependent to what extent these windows were further divided into smaller binning cells.

In order to adapt to the different frame rates of an action, two strategies were exploited. Window sizes were either readjusted to the general length of an action, with hard cutting boundaries, or fixed for a class, independent of the time of recording. Empirical investigation revealed best results, when the window size was fixed for class labels. The temporal extent considered is then consistent for all subjects of an action. Given that the frame rates effect the temporal dynamics of an action, window sizes were divided into three main subsets. This subdivision aimed at providing a rough adaption to the general dynamic differences of varying temporal expansions and speeds. The maximum number of frames for *run* was 179 and it was therefore given its own set. The next set defined the temporal extent for all actions except *run*, *salsa*, and *walk terrain*. Maximum time recording deflections for all actions in this subset ranged from 300 (*walk*) to 1200 frames (*dance*). *Salsa* and *walk terrain* were recorded over at least 2000 frames and therefore placed in the last set. The individual window sizes were initialized by the smallest set. Each set thereafter, either doubled the frames from the previous or added the same temporal extent of the *"running"* set on top. When the smallest window size considered 20 frames, for example, the others investigated either (40,80) or (40,60) frames.

The subsampling factor that creates smaller binning cells in each window was found to be stable and had an insignificant impact on performances. To avoid different feature vector lengths and hence time-series classifications this factor had to be kept consistent for all points of analysis. The four cells of each window originally selected performed equally well compared to smaller divisions. Since this factor consequenced reasonable computational complexity, these four cells were fixed throughout the course of the study.

Points along the diagonal

Analysis had to be restricted to a definite number of points in time and it was observed that a correct selection had large influence on the classification process. The number of analysis points on the diagonal defines to what extent the individual windows overlap. Finally, global parameter optimization was achieved by iterating through a collection of numbers and inspecting the performances for each setup. To add variability to the temporal extent considered at each point of analysis, it was similarly iterated over different window sizes. All pairs considered ranged from 2 - 25 points on the diagonal with 8 - 22 frames considered at each of these points for the smallest window (*run*).

	Average binning		All binning	
Parameters	Moving	Stationary	Moving	Stationary
Pixels on Diagonal	14	17	6	24
Window sizes	(16,32,64)	(20,40,80)	(21,42,63)	(20,40,60)
Performances (%)	90	86.7	86.7	70

Table 5.1. Classification performances and parameter selection summary for the average and all binning description.

Table 5.1 displays a summary of all parameter selections for both datasets and binning descriptions. Please note that there existed multiple setups that achieved identical performances. A generalized behavior and global settlement of performances for a certain parameter combination was not found. Moreover, Nearest Neighbor classifier ($k = 1$) achieved best results in all trials. The reasons for this were found to be very similar to the ones explained in the previous chapter.

5.3 Discussion

The principal goal of the chapter at hand is to inspect a recognition scheme for human motions, which describes the temporal dynamics of actions as absolute correlations of features over time. The corresponding dynamic patterns are analyzed using the HOG descriptor. Performances were coherently compared between subjects with available motion trajectories and stationary subjects. Analysis was extended for a feature description that utilized all gradient assignments against a global representation of averaged binning descriptions.

Results indicate an outperforming of the global description against a purely local representation. In the end, high across-view stabilities were observed for both moving and stationary samples, independent of the gradient description. Reducing the amount of training angles did not affect the obtained rates.

What follows targets to enlighten the view-independent usability of the proposed framework. The effect of removing global-person translation by normalizing the view as well as geometric rotations on entries in the self-similarity matrix is systematically explained. Furthermore, performance differences for spatial and moving samples are evaluated and reasoned. The difference between both gradient descriptions is afterwards assessed, based on the previously gained knowledge. In the end, it is striven to show that the presented approach is a promising action recognition framework, which has a hypersensitivity for parametric optimizations, but large extent of view-independent stability.

5.3.1 Isometries and view-independence

Being part of a group of rigid transformations, rotations and translation are isometries. Isometries preserve the distance relationships between all pairs of points of a rigid body with the transformation. What follows gives a short mathematical proof about the effect of rotations and translation on entries in the self-similarity matrix.

Rotations

Suppose that $\mathscr{R}_\theta : (z) \mapsto (e^{i\theta}\ z)$ denotes the rotation of an arbitrary body by an angle θ, since $e^{i\theta} = cos(\theta) + i\ sin(\theta)$. Let $a, b \in \mathbb{C}$

$$\begin{aligned} |\mathscr{R}_\theta(a) - \mathscr{R}_\theta(b)| &= |(e^{i\theta}a) - (e^{i\theta}b)| \\ &= |(e^{i\theta})|\ |(a-b)| \\ &= |(a-b)| \end{aligned} \tag{5.4}$$

and the distance between a and b, or all frame pairs, is the same as prior to the rotation. The resulting SSM is hence identical. Accordingly, the SSM has a definite view-independent property within angles of the same subject. This reasons why subjects are either entirely classified or misclassified under all angles of rotation and explains classification performances occurring in blocks of 33%, since only three subjects were represented in the test set.

Translations

Suppose that $\mathscr{T} : (z) \mapsto (z + c)$ denotes the translation of a point p by a coefficient c. Let $a, b \in \mathbb{C}$. It then holds that

$$\begin{aligned} |\mathscr{T}(a) - \mathscr{T}(b)| &= |(a + c) - (b + c)| \\ &= |a - b| \end{aligned} \tag{5.5}$$

and the distance between a and b is the same as prior to the translation. Opposite to rotating the skeleton, a translation results in different entries in the self-similarity matrix. This is due to the consistency of the rotation angle θ for rotations and to the varying transformation coefficient c for translations. Since the torso is subtracted frame-wise, c varies constantly. Translating the skeleton can then not be seen as a single transformation of one rigid body, but different translations at every frame. The distance relationship between pairs of frames with different coefficients therefore changes with each frame. This results in different entries in the SSM for the same subject after a translation.

Recall that HOG is a descriptor that solely resorts to local gradient orientations and directions. The complete classification process is based on a collection of these gradient binnings. Translations therefore do not distort the temporal dynamics of action representations themselves, just their superficial self-similarity structure, which is a consistent change in the view-normalization process. The parameters that achieved best performances for the spatial dataset, are apparently not transferable to the new structures given by the translation. They seem to overfit the underlying distribution and binning description and fail to generalize. A detailed interpretation of this sensitive parametric behavior is given next.

5.3.2 Feature detection and performances behavior

Problem of locality

As previously mentioned, analyzing the SSM was due to its symmetry restricted to the upper part of the diagonal. How many frames considered in each local inquiry determines the likelihood to capture the main characteristic structure of an action at that specific point in time. Different actions with diverging global gradient structures display close to equal local gradient directions, whilst same action classes, with similar global gradient structures, exhibit diverse structures if inspected under different point of views. This phenomenon is referred to as the problem of locality.

The window size defines the temporal extent considered. As a matter of fact, finding SSM characteristic structures depends on the length of recordings. For larger frame rates, characteristics spread along a larger temporal extent. It was observed that when the temporal extent investigated is too short, the probability that the structure is not adequately captured increases. A coherent adaption of the temporal scope was therefore examined. As previously mentioned, window sizes were either adapted to action lengths, or fixed for their labels.

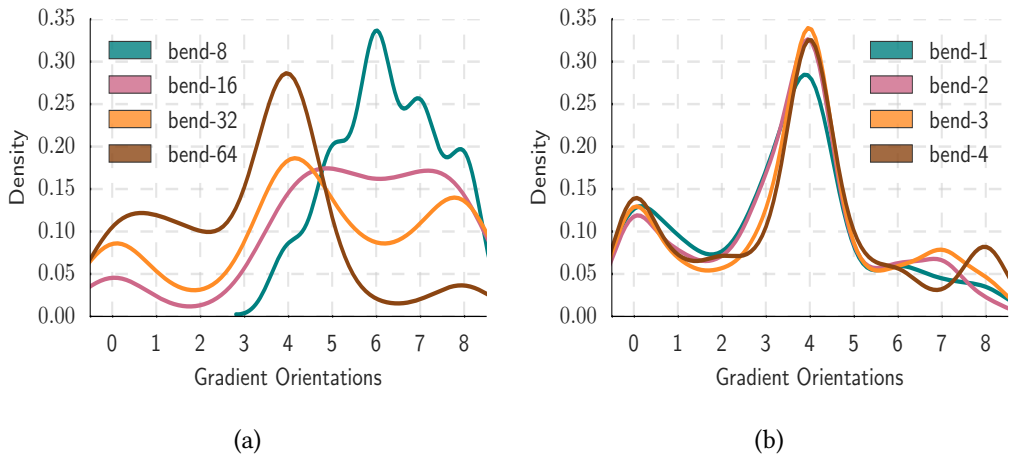


Figure 5.5. Kernel density estimation of maximum binnings for a bending action obtained after every iteration of HOG descriptor. The windows were either adapted to the length of an action, with temporal extents ranging from 8 - 64 frames (a), or the window size was fixed for the four actions (b). The label-fixed strategy creates a more stable binning description across subjects.

To the end, label-fixed boundaries outperformed an adaptive strategy. Figure 5.5 depicts the effect on the binning distribution for a bending action with an adaptive and fixed strategy. It was noted that binning behaviors change in such a distinct fashion for the adaptive strategy, that a coherent representation of multiple bending actions, where HOG was applied to at different scales, is difficult to obtain.

Label specific window sizes nonetheless display some solid weaknesses, when inter-class action recordings are vast. Literature pointed towards the difference in frame rates and the resulting different temporal speeds at which actions are performed (Dexter, Pérez, and Laptev, 2009). In some cases, fixing the temporal extent for class labels is not suitable

for capturing these speed fluctuations and results in concrete structural differences discovered by HOG. The advantageous window sizes for the action sets defined in this thesis were obtained by iterating over a large range of possibilities and hence more based on trial and error observations. Ultimately, no global optimization of window sizes was discovered, that would have indicated a generalized behavior for both spatial and view-normalized samples. Under these circumstances, it becomes necessary to find unique global action structures by either adapting the temporal analysis automatically at each iteration for motions with varying frame rates, or by fixing it for actions that were cut to unit length.

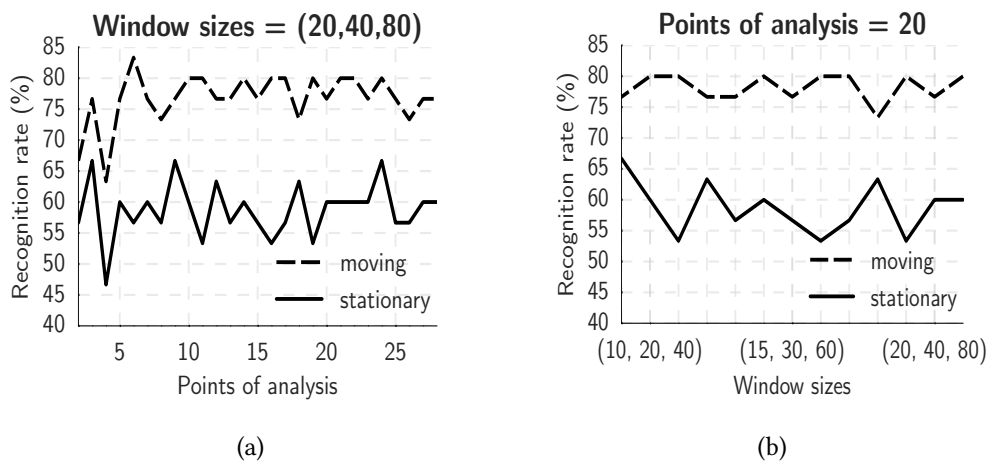
Dexter et al. (2009) introduce an automatic solution, by calculating the Laplacian of Gaussians (LoG) over a range of standard deviations at each iteration point on the diagonal. The corresponding scale space representation adapts to the differences in temporal speed and the scale maximizing the LoG is selected for the corresponding point of analysis. Instead of adapting the temporal extent at each point of analysis, the behaviors were inspected when all actions were manually cutted to unit length. This has the strong advantage of removing the need for window size adaptations, since the temporal expansion is equalized for all actions. It was observed that cutting resulted in large probabilities to miss action specific innate structures. Instead of finding dynamic or fixed boundaries for when an action starts or ends, the problem manifests itself within capturing the most informative, action-specific point of analysis. To the end, it was not possible to find satisfactory cutting strategies, that would have indicated a coherent representation for action labels. This is because small derivations in the global structure of the extracted parts have large effects on the local gradient computations of the HOG descriptor.

Points of analysis

Since the cutting strategies applied for this paradigm did not yield satisfying performances, hand in hand with a general avoidance of time-series classification, gradient computation had to be restricted to a definite number of points on the diagonal. This number was then kept consistent for all actions. Figure 5.6 depicts the evolution of performances when either the window size or the number of pixels on the diagonal was fixed. High performance fluctuations were detected for different number of points and window sizes, with no tendency for a global optimization behavior.

Further research could therefore investigate if an optimal global parameter setting can be found for spatial and view-normalized samples, when motions are aligned using dy-

namic programming, for instance, similar to Zhou and De la Torre (2012). Cutting these motions to unit length, whilst preserving the characteristic global structure would then create possibilities for a consistent temporal analysis of both structures and therefore remove the difficulty of window size adaptation.



(a) (b)

Figure 5.6. Performance evaluation for varying points of interest and window sizes for all binning feature vector. Fluctuations are detected when fixing the window size for varying number of points on the diagonal (a) and when the number of diagonal points are fixed for changing window sizes (b).

In summary, this resulted in the following observations. Adapting the detection window size for action labels ignores inter-class variations of actions, but was found to be more suitable than fixing it to the length of recordings. Since the temporal dynamics of actions are partially neglected in this case, it effects large performance fluctuations depending on the definition of the temporal scope. Similar behaviors were observed for the number of analysis points along the diagonal. As previously mentioned, the only difference between stationary and moving samples is the different analysis pattern inside the SSM. Nonetheless, it was detected that view-normalized samples performed worst for both gradient descriptions in all cases. The findings evaluated up to now, indicate that this is not due to the translation, but more a result of a hypersensitivity of the parameter space. Spatial samples outperformed stationary ones, given that the parameters were more suitable to represent the individual action structures.

5.3.3 Advantages of global representations

One of the main advantages of a HOG descriptor is an image representation entirely based on local shapes characteristic for that image (Dalal and Triggs, 2005). According to this knowledge, it was questioned why the global representation obtained through average binning outperformed classification based on local architectures solely. These findings are remarkably counter-intuitive to the general principles of feature detection.

Since averaging did not directly boost the performance of its all binning counter part with the same parametric setup, it had to be questioned why these representations resulted in superior performances. In the end, two possible causes were derived. This general improvement was either due to the more global representation scheme obtained through averaging, or due to the new feature description that fit the underlying distribution better out of coincidence. Despite the fact that no absolute solution was found for this controversy, it was hypothesized that averaging reduced the hypersensitivity of the parameter space. Murphy, Torralba, Eaton, and Freeman (2006) pointed towards the possible need for global image representations, when local information is insufficient, for small or partially occluded objects for instance. Interpreting the image as a whole can in this case partly remove the effect of missing local characteristics. It may therefore be possible that global binning representation provide better action descriptors.

It remains, nevertheless, open if an action description solely based on nine gradient orientations would suffice to distinguish larger datasets than the one used for this thesis. Especially for semantiy close actions it may be of importance to have all available local information as distinguishing factor. Global representations are therefore seen as a tool to moderately counter-act the problems that emerged from the parameter selection process. Adequate cutting solutions or coherent temporal adaption would probably remove its necessity.

Chapter 6

Conclusion

This thesis investigated the usability of two human action recognition frameworks. Of particular interest was to analyze their view-independent behavior and their sensitivity for view-normalization. A detailed overview of recent literature was given and the motion data introduced. The classifiers were presented thereafter. A precise assessment for both frameworks was provided in the end. What follows gives a comparative evaluation for both approaches by summarizing the main findings from before and highlighting their individual strengths and weaknesses for human motion understanding. An outlook for future lines of research will be given at last and close this chapter and thesis.

6.1 Summary and Comparison

The eigensubspace paradigm models human motions as a combination of average action features and dimensionality reducing projections. Classification performances were systematically compared between spatial and view-normalized samples under two different camera setups. This granted analysis of the general and view-independent usability. Investigation of the general usability with the all camera setup demonstrated robust performances up to 95.3%. Spatial information provided slightly better descriptors in this case. It was observed that across-view stability was guaranteed up to a certain minimal angular difference and correlated with a fast degradation of recognition rates thereafter. Spatial samples suffered from variance instabilities for complex motion categories and stationary samples from variance similarities for semantically related action classes. Removing spatial information when normalizing the view displayed greater stability, with peak performances of 87% overall.

The self-similarity paradigm analyzes the temporal dynamics of action features and models motions as changes in the structure of the dynamic action system. The changes are

described through characteristic peaks in the corresponding gradient orientations, and detected by the Histogram of Oriented Gradients descriptor. Analysis disclosed high across-view stability for rotations of the same subject and allowed the investigation of performances under a single training angle at 0 degrees. A degradation of performances was prevented due to the stability of distance relationships in the SSM with rotations. Performances were coherently compared between an average and all gradient binning description. A large view-independent stability was detected, once the structure of the self-similarity matrix was adequately captured. Best results were obtained for the averaged representation, ceiling at 90% across all angles. It was discovered, that the crux of classification is to optimize the parameter space of the HOG descriptor. The difficulty was to obtain coherent local representations for action classes recorded at different lengths and performed at varying speeds. Moreover, it was hypothesized that superior performances of the averaged representation are due to the more general gradient description, which reduces the hypersensitivity of parameter space. View-normalization changes the structure of the SSM and served to detect this sensitivity.

Evaluating the paradigms can be trimmed to certain quality criteria that allow a comparative evaluation of their practicability. A focus was appointed towards their view-independent behavior, computational complexity, necessity for parameter optimization and application spectrum. These criteria as well as general strengths and weaknesses are described for both frameworks next.

The computational complexity of the eigenspace approach exclusively relies on the linear transformations of the dimension reduction methods. In addition to the fixed feature vector length, this allows classification in real-time. The need for optimization is slight and at most depending on the parametrization of the dimension reduction methods. The approach does not require much initialization and the available features from the recording can directly be fed into the framework. Unfavorably, the paradigm requires specific joint locations as input and therefore suffers from being model-dependent. This limits the actual range of applications to systems with reliable joint capturing possibilities. Given that, performances dropped consistently from angular differences of 30 degree onwards, this thesis revealed a limited across-view usability.

The self-similarity descriptor on the other hand is beneficial were the eigensubspace has limitations and vice versa. Due to the large computational complexity of the HOG descriptor (Kim and Cho, 2014), analysis of the SSM pattern is remarkably time-consuming.

Scrutinizing the structure of the SSM was not achievable in real-time. Initializing the HOG descriptor such that it fits the underlying action best was very delicate to tune. This posits a main drawback compared to the naive initialization of the eigenspace approach. The framework shows promising view-stable properties, even for single views. The reasons for this is grounded within the definition of self-similarity representations themselves. Moreover, it works model free and analysis of temporal patterns does not require any specific type of action features.

In summary, it may be concluded that the self-similarity approach has a greater view-independent potentiality than that of the eigenspace approach. Temporal correlations seem to be less prone to view-fluctuations than variance information. Even though the parametric and computational complexity still posit some drawbacks, it has a larger leeway of analysis options and a wider range of applications due to its model-free functioning.

6.2 Future work and outlook

Future work could further investigate the usability of both paradigms by challenging the recognition task. It is known that most benchmark performances introduced by other studies are strongly dependent on the previous data selection process. Insight into this process is mostly not guaranteed. This thesis possessed for every tested subject its trained counter part with identical performer. This is beneficial for the classification process, since subject specific movements can previously be learned. Nevertheless, this does not represent a realistic recognition environment. Inspecting the behavior of the frameworks for a hard dataset, with entirely different performers in the training and testing set would be of crucial importance for future work.

Due to reasons of comparability and lack of available motion capture data for some classes, the dataset was kept consistent for both studies and limited to eight subjects per action. The second study revealed identical feature vectors for all rotation angles of a subject, which effectively resulted in three testing samples. Both approaches therefore need to be validated on larger datasets. Despite both of them are functioning considerably well, it has to be noted that they were only tested on the idealized MoCap environment. Körner and Denzler (2012) added "Salt-and-Pepper" noise to show the robustness of their approach. It remains open how the paradigm behaves for larger disturbances of the visual tracker, such as false joint localizations or complete missing of joints. The self-similarity approach has the advantage of being independent of these localizations. Junejo et al. (2011) have nonethe-

less only inspected their approach for the MoCap and KTH dataset. For the KTH dataset, features were manually initialized and placed on separate points of the body. It would be of interest to show its validity for recent benchmark datasets such as the Web-actions Dataset (Ikizler-Cinbis et al., 2009) or the Human Motion Database (HMDB51) (Kuehne, Jhuang, Garrote, Poggio, and Serre, 2011), which use selected motions from web videos that are non-idealized. Feature detection and classification is supposed to be much more difficult in these environments.

More research is needed for rotations along a second axis. It needs to be analyzed how different camera elevations affect the proposed schemes to model real-world scenarios. It would additionally be interesting to observe the behaviors for 2d data. The difference between the model-based eigenspace and model-free self-similarity approach may be of great importance in such a case, since stick-figure models are at the moment still difficult to obtain in 2d.

References

Aggarwal, J. K. and Quin Cai. Human motion analysis: A review. In *Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE*, pages 90–102. IEEE, 1997.

Akita, K. Image sequence analysis of real world human motion. *Pattern recognition*, 17(1): 73–83, 1984.

Bharatkumar, A., KE Daigle, MG Pandy, Qin Cai, and JK Aggarwal. Lower limb kinematics of human walking with the medial axis transformation. In *Motion of Non-Rigid and Articulated Objects, 1994., Proceedings of the 1994 IEEE Workshop on*, pages 70–76. IEEE, 1994.

Bobick, A. F. and James W. Davis. The recognition of human movement using temporal templates. *IEEE Transactions on pattern analysis and machine intelligence*, 23(3):257–267, 2001.

Boser, B. E., Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

Bottino, A., Matteo De Simone, and Aldo Laurentini. Recognizing human motion using eigensequences. 2007.

Bro, R., Evrim Acar, and Tamara G Kolda. Resolving the sign ambiguity in the singular value decomposition. *Journal of Chemometrics*, 22(2):135–140, 2008.

Cedras, C. and Mubarak Shah. A survey of motion analysis from moving light displays. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 214–221. IEEE, 1994.

Cortes, C. and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

References

Dalal, N. and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

Dexter, E., Patrick Pérez, and Ivan Laptev. Multi-view synchronization of human actions and dynamic scenes. In *BMVC*, pages 1–11. Citeseer, 2009.

Efros, A. A., Alexander C Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 726–733. IEEE, 2003.

Floyd, R. W. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

Hogg, D. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983.

Huang, P. S., Chris J Harris, and Mark S Nixon. Comparing different template features for recognizing people by their gait. 1998.

Ikizler-Cinbis, N., R Gokberk Cinbis, and Stan Sclaroff. Learning actions from the web. In *2009 IEEE 12th International Conference on Computer Vision*, pages 995–1002. IEEE, 2009.

Johansson, G. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14(2):201–211, 1973.

Junejo, I. N., Emilie Dexter, Ivan Laptev, and Patrick Perez. View-independent action recognition from temporal self-similarities. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):172–185, 2011.

Kazhdan, M. Shape representations and algorithms for 3d model retrieval. *Recommended for acceptance by the department of computer science*, 2004.

Kim, S. and Kyeongsoon Cho. Fast calculation of histogram of oriented gradient feature by removing redundancy in overlapping block. *J. Inf. Sci. Eng.*, 30(6):1719–1731, 2014.

Körner, M. and Joachim Denzler. Analyzing the subspaces obtained by dimensionality reduction for human action recognition from 3d data. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 130–135. IEEE, 2012.

Kuehne, H., Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

Li, W., Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 9–14. IEEE, 2010.

Lowe, D. G. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

Martens, J. and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1033–1040, 2011.

Marwan, N., M Carmen Romano, Marco Thiel, and Jürgen Kurths. Recurrence plots for the analysis of complex systems. *Physics reports*, 438(5):237–329, 2007.

Moeslund, T. B., Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2): 90–126, 2006.

Murase, H. and Rie Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern recognition letters*, 17(2):155–162, 1996.

Murphy, K., Antonio Torralba, Daniel Eaton, and William Freeman. Object detection and localization using local and global features. In *Toward Category-Level Object Recognition*, pages 382–400. Springer, 2006.

Poppe, R. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

Ramanan, D. and David A Forsyth. Automatic annotation of everyday movements. In *Advances in neural information processing systems*, page None, 2003.

Rao, C., Alper Yilmaz, and Mubarak Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

Schölkopf, B., Alexander Smola, and Klaus-Robert Müller. Kernel principal component analysis. In *International Conference on Artificial Neural Networks*, pages 583–588. Springer, 1997.

Shechtman, E. and Michal Irani. Space-time behavior based correlation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 405–412. IEEE, 2005.

References

Tenenbaum, J. B., Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Thurau, C. and Václav Hlavác. Pose primitive based human action recognition in videos or still images. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

Tipping, M. E. and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

Tran, K. N., Ioannis A Kakadiaris, and Shishir K Shah. Modeling motion of body parts for action recognition. In *BMVC*, volume 11, pages 1–12. Citeseer, 2011.

Wang, J., Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

Wang, J., Zicheng Liu, and Ying Wu. *Human Action Recognition with Depth Cameras*. Springer, 2014.

Wang, Q. A survey of visual analysis of human motion and its applications. *arXiv preprint arXiv:1608.00700*, 2016.

Weinland, D., Remi Ronfard, and Edmond Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *Computer vision and image understanding*, 115(2):224–241, 2011.

Zhou, F. and Fernando TorreDe la . Generalized time warping for multi-modal alignment of human motion. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1282–1289. IEEE, 2012.

Acknowledgments

First and foremost I wish to express my deepest gratitude to Pattreeya Tanisaro, for the continuous support throughout my thesis. Even though I was quite overwhelmed by the amount of input when I first started, she always supported me with her intense knowledge, patience and friendly attitude.

I would like to extend my gratitude to my closest friends and my family who guided me throughout this time. I owe special thanks to them for their continuous help. Without their encouragement I would not have been capable of weathering all the struggles that came up on the way.

Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other university.

October 26, 2016

Florian P. Mahner