Graz University of Technology

Institute for Computer Graphics and Vision

Master's Thesis

Human Action Recognition using Multiple Instance Learning: A Comparative Study

---

Fritz Gerald

Graz, Austria, Mai 2011

Thesis supervisors

Univ.-Prof. Dipl.-Ing. Dr. Horst Bischof

Dipl.-Ing Dr. Peter Roth

Deutsche Fassung:
Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008
Genehmigung des Senates am 1.12.2008

EIDESSTATTLICHE ERKLÄRUNG

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am …… ……..
(Unterschrift)

Englische Fassung:

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

…… ……..
date (signature)

To my parents ...

Abstract

This thesis is situated in the scope of human action recognition and is concerned with two major objectives. First, it presents a comparative study of five different multiple instance learning (MIL) approaches and relates the results to those reported for state-of-the-art approaches in this field. Second, this work considers whether a sparse, part-based representation is able to support the consecutive classification process.

We investigate a non-negative matrix factorization with sparseness constraints and determine how such a representation contribute to performance improvements. Furthermore, we analyse the impact of a structured initialization towards a better part-based representation and present results for two different nearest neighbour approaches in a face recognition experiment. In the main part of this thesis we investigate, whether a MIL concept is suitable for an action recognition task. We perform a thoroughly and detailed evaluation of different MIL approaches on the Weizmann action dataset and the KTH benchmark.

Results on the ORL database of faces demonstrate that sparse, part-based representation beneficially supports the subsequent classifier. In particular, if the level of sparseness is significantly greater than those obtained by an unconstrained matrix factorization, then both classifiers achieved an increased performance compared to the unconstrained feature representation. Results of our comparative study in HAR showed that three out of five MIL methods achieved competitive or better accuracies compared to a linear SVM classifier, when evaluated on the Weizmann dataset. Evaluations on the KTH benchmark demonstrate, that the best MIL approach (*miGraph*) performed equally well up to a moderate level of noise. Finally, a solid comparison with recent approaches in the field of human action recognition complements the discussion of both datasets.

Keywords. human action recognition, multiple instance learning, non-negative matrix factorization, sparse coding.

Kurzzusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der automatischen Erkennung von menschlichen Bewegungsmustern in Videosequenzen. Das hauptsächliche Ziel der Arbeit ist eine Vergleichsstudie von fünf unterschiedlichen Klassifikatoren, die alle im Kontext von 'multiple instance learning' (MIL) definiert sind. Ein weiteres Thema dieser Arbeit ist die Problematik der Merkmalsrepräsentation. Hier wird untersucht, in wieweit eine zusätzliche Limitierung der Merkmale auf Objektteile den nachfolgenden Klassifikationsprozess unterstützen kann.

Im Speziellen wird eine Variante der 'non-negative matrix factorization' (NMF) untersucht, die eine weitere Restriktion in der Repräsentation erlaubt. Zusätzlich wird der Einfluss einer strukturieren Initialisierung dieser Faktorisierung analysiert. In der Folge werden zwei unterschiedliche Klassifikatoren ('k-nächste Nachbarn Algorithmus') evaluiert und deren Resultate im Kontext der Gesichtserkennung verglichen. Im Hauptteil dieses Werks erfolgt eine intensive Auseinandersetzung mit MIL, die der Frage nachgeht, ob ein solches Konzept für die Problemstellung der Erkennung von Bewegungsmustern geeignet ist.

Im Bereich der Gesichserkennung konnte gezeigt werden, das sich die alternative Repräsentation von Merkmalen positiv auf das Klassifikationsergebnis auswirkt. Eine geeignete Parametrisierung des Faktorisierungsprozesses erlaubte eine höhere Erkennungsrate im Vergleich zu der Standardmethode. Die Ergebnisse der Studie im Bereich menschliche Bewegungsmuster zeigten, dass drei von fünf MIL Ansätzen bessere oder zumindest gleichwertige Resultate erzielten, wenn sie mit einer herkömmlichen Klassifikationsmethode verglichen werden. Die Auswertung auf einer umfangreicheren Datenbank lassen den Schluss zu, dass zumindest ein MIL Algorithmus (*miGraph*) gleichwertig performant ist, sofern das Rauschen einen moderaten Anteil nicht übersteigt. Ein umfangreicher Vergleich mit anderen Arbeiten im thematischen Umfeld rundet die Evaluierung auf der jeweiligen Datenbank ab.

Acknowledgements

First of all, I would like to thank my thesis supervisor Univ.-Prof. Dr. Horst Bischof, for the opportunity to finish the work, for his ideas, his patience and for asking 'Is there room for hope?'

I am grateful for having had the opportunity to be advised by Dr. Peter Roth. Many thank for your guidance, for focussing my attention to details and for contributing many alternative viewpoints. This thesis has been greatly improved by your valuable feedback.

I further would like to thank Thomas Mauthner for providing me with the pre-calculated features of both datasets used in this work, Alexander Moschig, for his friendship, for proof reading several sections of the thesis and for helping me in LaTeX situations.

I would like to thank my parents for raising me and their support during my whole life.

Finally, but most importantly, I would like to thank you, Sandra, for your love and appreciation. Without your solid foundation of support all the way through this project, the thesis would never have been completed.

Graz, May 16 Gerald Fritz

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 State of the Art in HAR 3

1.3 Objectives and Outline of the Thesis 8

2 Sparse Representation **10**

2.1 Part-based and Sparse - Two Principles 10

2.2 Part-Based Representation via Sparse NMF 12

2.3 Experiments 15

2.3.1 ORL Database of Faces 15

2.3.2 Sparse NMF using Structured Initialization 17

2.3.3 Recognition Results 20

2.4 Conclusions 21

3 Multiple Instance Learning for Human Action Recognition **23**

3.1 Multiple Instance Learning 24

3.1.1 Supervised vs. Multiple Instance Learning 25

3.1.2 Formalization of the MIL concept 26

3.2 MIL Algorithm Reviewed 27

3.2.1 Kinematic-MIL 28

3.2.2 miGraph 30

3.2.3 MILES 31

3.2.4 MI-SVM 33

3.2.5 Citation k-NN . . . 34
3.3 Comparative Study . . . 36
3.3.1 Histogram based feature representation . . . 37
3.3.2 Datasets for Human Action Recognition . . . 39
3.3.3 Evaluation Method . . . 42
3.4 Experimental Results . . . 44
3.4.1 Static and dynamic features for Action Recognition . . . 45
3.4.2 Performance on the Weizmann Dataset . . . 46
3.4.3 Comparison to previous papers . . . 49
3.4.4 Performance on the KTH Dataset . . . 50
3.4.5 Robustness to Noisy Patches . . . 52
3.4.6 Comparison to previous papers . . . 55

4 Conclusion **58**
4.1 Summary . . . 58
4.2 Summary of our Contribution . . . 60

Bibliography **61**

List of Figures

2.1 Schematic illustration of the LMNN approach 15
2.2 ORL Database of Faces 16
2.3 NMF basis images from the ORL dataset 18
2.4 Sparse NMF: random vs. structured initialization 18
2.5 Approximation error: random vs. structured initialization 19
2.6 Recognition performance using sparse features 21

3.1 Supervised vs. Multiple Instance Learning 26
3.2 Key Chain Example 26
3.3 Instances as non-I.I.D. Samples 31
3.4 Citation k-NN Example 36
3.5 Overview Weizmann dataset 40
3.6 Overview KTH action dataset 42
3.7 Confusion Matrix original vs. mirrored dataset 47
3.8 Confusion matrix for miGraph on the KTH action dataset 53
3.9 Schematic illustration of simulated detector noise 54

List of Tables

2.1 Average sparseness achieved using NMF 17

3.1 Summary of notation used for MIL 28
3.2 Improvement using HOG and HOF features 46
3.3 Performance for the Weizmann action dataset 48
3.4 Results for the Weizmann nine action subset benchmark 49
3.5 Weizmann Recognition Precision Ranking 50
3.6 Performance comparison on the KTH benchmark 51
3.7 Performance evaluation with different noise levels for the KTH 54
3.8 KTH Recognition Precision Ranking 56

If you would be a real seeker after truth, it is necessary that at least once in your life you doubt, as far as possible, all things.

Rene Descartes

Chapter 1

Introduction

1.1 Motivation

This thesis is situated in the scope of human action recognition (HAR), an emerging research direction in the area of computer vision during the last decade. The interest in this area is motivated from different fields of application. The fast growing amount of on-line videos for example, raises new challenges in the context of semantic search. The increasing number of surveillance cameras in a broader area of every day live promise a wide field of applications. While manual inspection of this data become more and more impractical, efficient methods for automatically recognizing unexpected action could help us to observe critical areas at lower costs and could raise public safety. Similar methods can be applied to sport videos, where the analysis of strategies and tactical movements become a typical business for trainers and athletes.

It is important to thoroughly define the context of HAR, to distinguish our work from similar research directives. Since recognition of actions can be performed at different levels of abstraction, Moeslund et al. [46] introduced a taxonomy on that issue. Their hierarchy established action primitives, actions and activities as basic concepts. The authors defined action primitives as simple gestures at limb level, actions as cyclic or single whole-body movements, and activities as a number of subsequent actions. Following these definitions, we concentrate on actions as single body movements and do not consider interactions between persons [56] or objects. Furthermore, the objective of HAR is contrary to the field of gait recognition. Opposite to the goal of identifying personal styles of e.g. walking, we have to generalize across these variations. Finally, it is important to mention that we assume full-body movements, i.e., the whole human body has to be visible in the video

data, and do not consider partial occlusions.

The process of feature extraction and representation is significant in every recognition framework. First, the procedure has to preserve the relevant information from the visual environment and, second, the encoding has to be sufficiently rich to allow for robust classification. Due to the holistic nature of our feature extraction and the fact that complex actions could be composed of multiple limb movements, we hypothesize that a intermediate part-based representation of these features would support the later classification process. In particular, we consider an algorithm called non-negative matrix factorization (NMF), that allows only additive combination of intermediate parts, as one method to achieve this kind of representations.

Machine learning concepts stimulated the field of computer vision over many years. Discriminative classifiers like Support Vector Machines (SVM) have been used by many authors [28, 33, 43, 47, 58] as part of their action recognition framework. Another popular method for supervised learning is k-nearest-neighbour (k-NN) classification. This simple but fundamental algorithm was used by e.g. [18, 38]. More sophisticated approaches, like cascaded Linear Discriminant Analysis classifier (LDA) used by Roth et al. [55] or probabilistic models used by Niebles et al. [48], demonstrate the variety of machine learning methods in the area of HAR.

When applying supervised learning techniques, one decisive question emerges: How much annotation is needed? The assumption of perfect labelled data results in limitations for many proposed algorithm, when dealing with real world applications. For example, manual annotation of many hours of surveillance videos is tedious and expensive. One possibility to overcome this limit is Multiple Instance Learning (MIL). Inspired by the recent advances in the field of face detection [5] and visual tracking [6], we strongly suspect that this learning concept fits well into the context of HAR. The application of MIL in this context is motivated in the following.

In the concept of MIL, samples (denoted as *instances*) are organized in *bags* and a bag is simply a set of instances, which share a common label. In the scope of action recognition a bag comprises all frames from a video presenting a specific action. Moreover, each frame itself is represented by multiple image regions, which describe the locations of a single or multiple persons performing that action. Hence, a instance or sample denotes the encoding of an region showing the human body performing an specific action, and this instance is extracted from a single frame of the video. The MIL concept assumes a bag being positively labelled, if it contains at least one instance describing the specific action,

whereas a negative bag contains only instances describing other actions.

These samples could be extracted manually or automatically using for instance the method proposed by Dalal & Triggs [13]. As a consequence of an automatic process, these samples might be misaligned with the human body and false detections may occur. To create a perfectly labelled training dataset for supervised learning, one would need to either manually localize the human body in every frame or filter out imperfect detection results. In addition, each sample has to be labelled according to the performed action. Differently in case of MIL, where only bags are annotated. Moreover, based on the MIL assumption, a bag labelled positive (e.g., for action walking) comprises instances representing a walking action as well as negative instances, e.g., those caused by an imperfect detector. The objective in MIL is to derive a classifier that deals with the ambiguous instances within the labelled trainings bags.

Negative instances originated by false detections are one argument for using MIL, imprecise temporal segmentation of actions is another one. It is difficult do define the exact beginning and ending of an action within a continuing video stream. Nevertheless, MIL is able to handle this issue, due to the MIL assumption. Additional frames from the beginning or end will generate further negative samples within the bag. However, under the assumption that the positive action is performed between start and end frame, there will be at least one positive instance. Therefore, the label of the bag will be positive for the particular action.

Consequently, the effort of manual annotation could be heavily reduced using automatic person detection methods, and the labelling could be performed based on sequence labels, in contrast to a frame based interpretation.

The next section summarises recent approaches in the scope of HAR, without claiming completeness. The reader is referred to Poppe [51] for a more comprehensive overview on vision based human motion analysis.

1.2 State of the Art in HAR

Similar to the task of object recognition in single images, where a vast number of feature detectors and descriptors have been developed [44], action recognition caused different novel techniques for detecting and describing local portions in space and time, often called spatio-temporal volumes.

Efros et al. [18] for example proposed an approach on action recognition several years ago. Motivated from broadcast videos, where humans are typical around 30 pixels tall,

they introduced a novel motion descriptor for recognizing actions at a distance. For that they computed spatio-temporal volumes of each individual person by tracking them from frame to frame. The representation is based on optical flow images in x and y direction and since these are known to be noisy, their blurred versions are used for later recognition in order to deal with small misalignments in space and time.

Alike traditional interest point detectors Laptev and Lindeberg [32] extended the notion of spatial interest points to a temporal dimension. They introduced a detector that is sensitive to pixel values with significant local variations in space and time. Their experiments demonstrated that such locations often correspond to interesting events in the video. By clustering local volumes around these events they evaluated a single action model for a walking person and where able to detect and estimate the pose in the test videos.

A novel approach to extract spatio-temporal features was proposed by Dollar et al. [15]. As an alternative to space-time interest points [32], their detector is sensitive to local intensity variations in the image that contain periodic frequency components. The spatio-temporal neighbourhood of the detected location was used to compute different descriptors, where a flat representation of image gradients in combination with a *Principle Component Analysis* (PCA) worked best. This method was used later on by the same author [33] and others [47] within their frameworks.

Other approaches pursued a dense sampling approach, i.e., Schindler and van Gool [58]. They used form features based on Garbor filter banks and a dense optical flow representation to encode the discriminative patterns in various response maps. Their classifier, a combination of linear SVM, evaluates unseen videos frame-by-frame. The authors claimed that a small portion of the action video is sufficient to achieve robust and accurate recognition results. Interestingly they found out that an additional weighting between appearance and motion features increases the overall accuracy. However, experiments showed contradictory results, when comparing results from two state-of-the-art benchmark datasets in action recognition. This leads us to the conclusion that no general valid statement could be made. In section 3.4.1 we describe our results on an experiment investigating the influence of static and dynamic feature combinations.

Junejo et al. [28] proposed an alternative concept to overcome several limitations of current methods like simple background, static camera, or limited view variations. They considered action recognition under view changes and developed a representation of actions that captures structure of temporal similarity and dissimilarity. These descriptors

are extracted from the self-similarity matrix of different patch based or body part trajectory representations of human actions. Neither multi-view correspondence estimation nor explicit modelling of body parts is needed by their concept. Although designed for recognition across multiple viewpoints they showed state of the art performance results on the single view Weizmann database (see section 3.3.2 for a description of this dataset). However, their approach need a sufficient large number of frames per video to extract a useful similarity matrix and therefore processing the whole video data is usually required.

Ali et al. proposed a new kind of kinematic features set and showed improved performance by using MIL in an human action recognition task [2]. The contribution of their work is twofold. First they developed a new set of features that are derived from the optical flow of a sequence of images. They call them *kinematic* features motivated by the assumption that different aspects of dynamic motion patterns within the optical flow field can be represented more discriminative compared to the standard flow. Second and more related to our work is their computation of *kinematic modes* and the application of multiple instance-based learning in the context of human action recognition. Following the spirit of the diverse density framework originated by [14] they projected their feature representation into a single instance space and a k-NN classifier was trained to predict the actions in unseen videos.

It is obvious that the applied representation has to be sufficiently rich to allow for robust recognition of the action. It also seems to be accepted in the community that a combination of multiple cues (e.g. appearance and motion) beneficially supports the later classification. The reader is referred to the work of Wang et al. [65] for a comprehensive study on valuable detector/descriptor combinations.

One study that clearly showed the benefit of appearance and motion features was presented by Mauthner et al. [43]. Their framework for classifying short action sequences is able to recognize human actions by using only two frames. They showed that similar poses present in different actions lead to an ambiguous representation in the appearance cue. These actions can be distinguished by the additional use of motion information. Therefore, a histogram based representation of both pixel intensity and dense optical flow features was used and further processed by applying a NMF to get compact and robust features. Multiple SVMs complete their frame-wise action recognition system and during experiments they showed highly competitive performance results on the Weizmann database.

Contrary to the previous work Thurau and Hlaváč [62] ignored motion features at all. They concentrated on the idea of view or pose based action recognition from single images and image sequences. Assuming that different actions share common pose appearances, their approach represented different actions as histograms of pose primitives. These prototypical pose descriptors are learned from training data by separating them from clutter via simple background subtraction. During training NMF is used to build a lower dimensional representation of both human poses and background. Agglomerative clustering reduced the number of prototypical descriptors and an additional information criteria introduced a weighting for more informative cluster entities w.r.t particular actions. They showed high competitive recognition performance on the Weizmann dataset when dealing with single- and multi-frame evaluation.

While the aforementioned methods consider action recognition and assume a separate preprocessing step for localizing them, more advanced approaches perform both objectives simultaneously.

Niebles et al. [48] presented an unsupervised learning method that is able to recognize and localize multiple actions. Based on space-time interest points [15] they used a k-means clustering with a Euclidean distance metric to derive a code book representation. The authors empirically evaluated two probabilistic models and showed their robustness against noisy features. Although Niebles et al. claimed their method to be unsupervised in the sense of action classes, the number of different actions has to be known in advance. However, their algorithm is able to recognize and localize multiple actions, even in long video sequences.

A different method aiming recognition and localization was introduced by Mikolajszyk and Uemura [45]. They learned a vocabular forest by extensive feature extraction out of the trainings data using different detectors and features describing gradient, edge, as well as motion information. The authors used a model-based approach to represent actions in order to encode global structure and the relations among moving parts. The major contribution of their work was a model based action representation that makes their approach able to localize multiple actions in the image, while a robust and efficient processing was achieved by the use of multiple vocabular trees.

A recent approach for HAR was proposed by Yao et al. [72]. Their Hough transform-based voting method is able to classify and localize actions in a probabilistic framework. Yao et al. trained one single classifier, while sharing common features across multiple action classes. Their representation is a composition of low-level features derived from

gradients and optical flow estimates. By utilizing randomized trees and following a dense sampling strategy they achieved state-of-the-art performances on different standard benchmarks including UCF sports dataset [53] and several surveillance scenarios from the UCR Videoweb activities dataset [1].

Another bag-of-word approach was proposed by Laptev et al. [33]. They addressed the problem of more realistic HAR. In particular, they considered an automatic annotation and segmentation of the video data by the use of movie scripts and subtitle information in combination with automatic text classification algorithms. They used a combination of shape and motion features to describe the location around previous detected space-time interest points. Like other authors [38, 48] they build a bag-of-word representation for both HOG and HOF features and evaluated various grid configurations for feature extraction in the space time volumes. They demonstrated the robustness of their method with 10 % decrease in accuracy in the presence of 40% label noise on the KTH dataset and other more challenging benchmark datasets. However, their automated video annotation algorithm is limited to scenarios where scripts and subtitles are available.

Niebles and Fei-Fei [47] proposed a hierarchical approach that combines static shape features and motion features based on spatio-temporal interest points [15] on the lower layer and proposed a constellation model by combining parts on an intermediate layer of their hierarchical model. Exploring the contribution of different feature types they empirically showed that on the Weizmann dataset the combination of both features provides the best representation. In addition, the motion cue is preferable when using a single feature type.

Recently, another hierarchical framework was presented by Kovashka and Graumann [31]. They extended the bag-of-word representation of videos and proposed a novel compound feature-centred descriptor that is build on multiple levels of their hierarchy. A specific neighbourhood formation in space and time is used to generate the compound feature on the next upper level. The initial descriptors are conducted either from dense interest points and their HoG3D features [30] or from HOG [13] and HOF [43] representations for a sparse sampling strategy. They evaluated their framework on the UCF sports benchmark as well as on the KTH dataset of actions. Section 3.3.2 provides a brief description of the KTH data.

Special focus has been pointed towards efficient methods by the promise of interesting application areas. Like Mikolajszyk and Uemura [45], Roth et al. [55] proposed a very

[1] `http://vwdata.ee.ucr.edu`

efficient method for recognizing human actions. While using the same representation as [43] to guarantee effective and real-time feature extraction, an efficient cascaded Linear Discriminant Analysis (LDA) classifier significantly speeds up the overall performance of their framework compared to SVM based approaches. In addition, they empirically demonstrated that a weighting between different cues such as motion and appearance does not help to increase the performance and refuted suggestions by other authors [58]. Roth et al. achieved state-of-the-art performance results on standard benchmark datasets and showed that their framework is able to detect human actions on more challenging data in real-time.

Mauthner et al. presented a novel prototype-based action recognition framework in [42]. A sufficiently rich representation based on multiple feature cues was used to derive a vocabular tree for each feature channel individually. To further increase the classification performance, they proposed a temporal weighting of cues that depends on the current input data. Since this weighting is estimated during training, their method allows an efficient action recognition on a short-frame basis. A comprehensive evaluation was performed on the KTH and the Weizmann dataset and their approach outperformed existing state-of-the-art methods.

1.3 Objectives and Outline of the Thesis

We summarized the main objectives of our work in the following.

- We consider whether a sparse part-based representation is able to support the subsequent classification stage. We investigate a constrained NMF approach as well as a structured initialization method to obtain such a representation. We evaluated the proposed feature encoding in the context of face recognition and in an action recognition framework.
- The major objective is a comprehensive analysis of five different MIL approaches in the scope of HAR. We performed a thoroughly evaluation using two state-of-the-art benchmark datasets and compared our results with those achieved by other authors.

The remainder of this thesis is structured as follows:

Chapter 2 concerns the issue of sparse part-based representation, discusses the NMF with sparseness constraints and the particular choice of encoding parameters. It presents results of two nearest neighbour classifiers and demonstrates the advantages of the particular feature encoding regarding recognition performance.

Chapter 3 introduces the basic concept of MIL, reviews the investigated MIL approaches, describes the evaluation schema of our comparative study and presents results on two different action recognition benchmarks.

Chapter 4 provides a brief summary of this work and reviews our contributions.

Reality is only a Rorschach ink-blot, you know.

Alan Watts

Chapter 2

Sparse Representation

Contents

2.1 Part-based and Sparse - Two Principles 10
2.2 Part-Based Representation via Sparse NMF 12
2.3 Experiments 15
2.4 Conclusions 21

The major objective of this chapter is to empirically demonstrate that a sparse part-based representation supports the consecutive classifier. Hence, we evaluated the sparse *non-negative matrix factorization* (NMF) approach [26] in a face recognition task, that was designed as an intermediate evaluation step towards our main topic – HAR. Although, any other recognition task would have been appropriate, we focussed on this restricted problem domain, to concentrate on an extensive exploration of the parameter space. This would have been costly with the datasets used in section 3.3.2. Furthermore, the visualization of e.g., basis faces allows an intuitive interpretation and the effect of different sparseness constraints can be easily illustrated. This would have been more complicated in case of HAR, since the histogram-based features decompose the geometric information of the subjacent image. However, we used the methods and findings described here to encode the features for our comparative study in the context of HAR (see chapter 3).

2.1 Part-based and Sparse - Two Principles

Part-based representations are motivated by psychological and physiological findings, and several theoretical approaches for object recognition are drawn upon such theories [7,

17]. Lee and Seung demonstrated that NMF is able to learn a part-based representation in the context of face recognition [34]. The same authors [35] as well as Paatero [50] proposed different algorithms for this factor analysis technique, however, the multiplicative algorithms by Lee and Seung have been widely used for various problems in computer vision due to their simplicity. Guillamet et al. [23] compared the representations obtained by NMF and Principal Component Analysis (PCA). They empirically demonstrated the advantages of NMF in a patch classification framework and showed better result compared to PCA, especially, for scattered classes. Furthermore, NMF was used to classify faces in [24], but the technique of factor analysis has been used in other research areas as well. Shahnaz et al. [60] for example introduced a method to identify and cluster semantic textual features within documents.

A few years ago sparse representations have become popular in the area of signal processing and information theory. Donoho provided an abstract framework on this issue in [16]. He showed that the combinatorial problem of finding sparse solutions to systems of linear equations can be solved by minimizing the L_1 norm. Wright et al. [71] used these findings and proposed a robust method for face recognition in the presence of occlusion. Their algorithm encodes the test image as sparse linear combination of the trainings data and models the sparse error caused by the occlusion. They achieved superior performance results using raw images without any dimension reduction or feature selection.

Another interesting approach was proposed by Agarwal et al. [1]. The authors focussed, in contrast to Wright et al., on an object detection task, i.e., distinguish between object and non-object images and additionally localize the objects present in the image. They performed feature extraction on interest points and automatically constructed a vocabulary based representation for each object. The learned classifier used information regarding the spatial relations among parts and were evaluated on challenging natural imagery.

The combination of these two principles, the part-based representation derived via NMF and a sparse encoding, was exploited by Heiler and Schnörr [25] and Hoyer [26]. Both approaches extended the basic NMF formulation by introducing sparseness constraints and derived different algorithms to solve the modified problem statement. Heiler and Schnörr defined the sparsity-constrained NMF problem by introducing sparseness intervals, that have to be predefined by the user. They developed efficient algorithms based on second order cone programs (SOCP), which could be solved efficient and fast. The performance evaluation demonstrated that their methods perform similar compared to

the multiplicative update algorithm by Lee and Seung, with faster convergence and higher accuracy for high-dimensional datasets. They achieved similar reconstruction errors with less computational effort compared to Hoyer's approach.

However, in our work we focused on the work of Hoyer [26] due to two major reasons. First, the sparseness could be accurately controlled for basis vectors and coefficients using one value for each. This confines the number of parameters, which have to be optimized during training. Second, both methods perform almost equally well w.r.t the reconstruction error, while [25] is the faster algorithm. Since our evaluation focused on comparing MIL algorithm, we decided to accept the increased computational effort in the representation stage of our framework. The final and more practical reason was that the algorithm of Heiler and Schnörr depend on a commercial SOCP solver [1].

2.2 Part-Based Representation via Sparse NMF

This section introduces the basic concepts and methods used in our experiments. Starting with the multiplicative algorithm for NMF proposed by Lee and Seung [35], we give a brief review of Hoyer's sparse NMF extension [26], present a structured initialization approach for NMF and conclude with a brief overview of the two nearest neighbour classifiers used in Section 2.3.

Non-negative Matrix Factorization: The problem statement of a NMF can be written as follows. Given a $m \times n$ non-negative matrix $\mathbf{V}$, find two non-negative factor matrices, $\mathbf{W}$ and $\mathbf{H}$ such that

$$\mathbf{V} \approx \mathbf{WH}, \tag{2.1}$$

where $\mathbf{W}$ is a $m \times r$ matrix of basis vectors and $\mathbf{H}$ is the $r \times n$ matrix of coefficients. Usually, r is chosen such that $r \ll n$, to reduce the dimensionality of the feature space. NMF constrains all non-zero elements in $\mathbf{W}$ and $\mathbf{H}$ to be positive and every column vector $\mathbf{v}$ in $\mathbf{V}$ can be written as $\mathbf{v} \approx \mathbf{Wh}$. This is an approximation by a linear combination of all basis vectors, but contrary to PCA, this combination is purely additive. We used the multiplicative update rule and the squared error function (euclidean distance) as proposed in [35] as a baseline for our evaluation. The resulting optimization problem can be

[1] MOSEK: `http://www.mosek.com`

described as

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & \| \mathbf{V} - \mathbf{WH} \|^2 \\ \text{s.t.} \quad & 0 \leq \mathbf{W}, \mathbf{H}. \end{aligned} \tag{2.2}$$

This problem is convex in either $\mathbf{W}$ or $\mathbf{H}$, therefore only local optimal solutions are guaranteed.

NMF with Sparseness Constraints: Since our objective is to investigate sparse part-based representations, we used the approach proposed by Hoyer [26]. Contrary to Hoyer, who has only demonstrated that his algorithm is able to derive a part-based representation, we have investigated whether such an representation can increase the performance of the consecutive classifier. In the following we give a brief overview of this algorithm.

Hoyer's NMF extension allows the user to control the sparseness of $\mathbf{W}$ and $\mathbf{H}$ explicitly. The sparseness measurement used in [26] is defined as

$$\text{sparseness}(\mathbf{x}) = \frac{1}{\sqrt{n} - 1} \left(\sqrt{n} - \frac{\| x \|_1}{\| x \|_2} \right), \tag{2.3}$$

where n denotes the dimensionality of $\mathbf{x}$. This measure, which is based on the L_1 and L_2 norm, evaluates to zero, if all components of $\mathbf{x}$ have identical magnitude. On the other hand, a value of one can only be attained, if and only if exactly one non-zero component exists in $\mathbf{x}$. The constrained NMF problem can therefore formulated as follows:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & \| \mathbf{V} - \mathbf{WH} \|^2 \\ \text{s.t.} \quad & 0 \leq \mathbf{W}, \mathbf{H} \\ & \text{sparseness}\,(\mathbf{w}_i) \;=\; S_w \\ & \text{sparseness}\,(\mathbf{h}_i) \;=\; S_h, \end{aligned} \tag{2.4}$$

where $\mathbf{w}_i$ denotes the i^{th} column of $\mathbf{W}$ (i.e., a basis vector) and $\mathbf{h}_i$ denotes the i^{th} row of $\mathbf{H}$. Hoyer's NMF algorithm is a projected gradient descent algorithm that utilizes a projection operator, which guarantees the sparseness constraint for all columns of $\mathbf{W}$ and rows of $\mathbf{H}$ during each iteration.

NMF Initialization: We already mentioned previously that NMF may converge to a local solution. It heavily depends on the starting conditions, to which solution the NMF converges. The initial estimates for $\mathbf{W}$ and $\mathbf{H}$ affect the final solution and the number of iterations till convergence. Most applications initialize both matrices with random positive

numbers. However, we show in section 2.3 that a more structured initialization leads to a better part-based representation.

In our framework the Non-Negative Double Singular Value Decomposition (NNDSVD) schema was used to avoid the initialization dilemma. This method, proposed by Boutsidis and Gallopoulos [9], exploits the SVD as a optimal rank-r approximation of the matrix $\mathbf{V}$, where r denotes the number of basis vectors (also known as modes). The algorithm computes the r leading singular triplets of matrix $\mathbf{V} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. In each iteration step $j = 1 \dots r$, a matrix $\mathbf{C}^j = \mathbf{u}_j \mathbf{v}_j^\top$ is obtained and its non-negative section $\mathbf{C}_+^j$ is used to form the maximum singular triplet $(\mathbf{u}, \mathbf{s}, \mathbf{v})$. The column $\mathbf{w}_j$ of $\mathbf{W}$ is then initialized to $\mathbf{u}$ and the row $\mathbf{h}_j$ of $\mathbf{H}$ is set to $\mathbf{v}^\top$ scaled by the singular value s.

The NNDSVD provides initial values that enables the subsequent NMF algorithm to reduce the initial residuals after very few iterations [9]. Furthermore, the algorithm is able to exploit some of the inherent structure of the data. Primarily, the deterministic nature of NNDSVD provides the follow-up NMF with a static initialization, i.e., multiple runs will converge to the same local minima.

Nearest Neighbour Classifier: This chapter was intended to demonstrate that a sparse representation supports the consecutive classifier. Therefore, we evaluated the performance of a nearest neighbour (NN) classifier using these features.

Every classifier is based on some similarity measurement and in the absence of prior knowledge, the Euclidean distance is a widely used metric for that purpose. However, we want to emphasize that the basis vectors of NMF are non-orthonormal, contrary to those derived by PCA. The additional constraints regarding non-negativity cannot yield orthonormal bases.

To compensate for that, we evaluated an additional classifier approach and made sure that our results aren't affected by this issue. The *Large Margin Nearest Neighbour* (LMNN) classifier was introduced by Weinberger et al. [69]. They proposed a method for learning a Mahanalobis metric from labelled examples in a NN classification framework. Figure 2.1 illustrates the schematic overview of this method. The left side shows the local neighbourhood of sample $\mathbf{x}_i$ based on the Euclidean distance before training. The objective during learning is to optimize a metric, such that k-nearest neighbours of $\mathbf{x}_i$ always share the same class label. Simultaneously the metric has to separate examples from different classes (blue and red samples) by a large margin. The result of the trainings step is illustrated on the right side of fig. 2.1. Neighbours of $\mathbf{x}_i$ with similar label lie close to $\mathbf{x}_i$, while differently labelled samples are separated by a large margin. Weinberger et

al. proposed an efficient algorithm for solving such an optimization problem by using a semidefinite program and demonstrated the approach in the context of face recognition, text categorization, recognizing handwritten digits, and others.

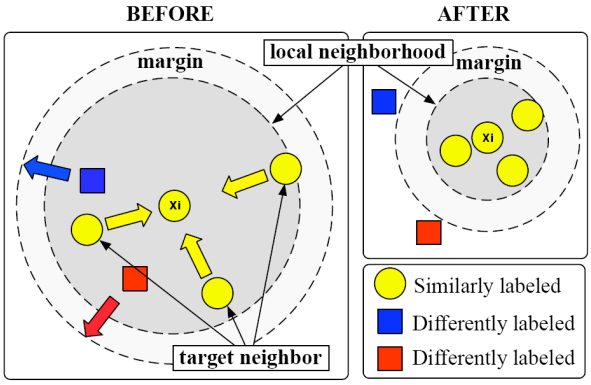


Figure 2.1: Schematic illustration of the LMNN approach: The similarity metric is optimized such that neighbours sharing the same label lie with a smaller ϵ-environment after training. Differently labelled samples are separated via a margin, which has to be at least the distance from $\mathbf{x}_i$ to its similar neighbours. Reprinted from [69].

2.3 Experiments

The experiments in this section are designed to demonstrate that a sparse part-based representation could support the subsequent classifier. Therefore, we focussed on a restricted problem domain — face recognition. Although, any other recognition task would have been appropriate, the encoding of faces allows an intuitive interpretation of intermediate results, e.g., basis faces, which could be easily visualized. The task involved the recognition of a particular face from an unseen pose. During training we derived different representations using unconstrained and sparse NMF with structured initialization (NNDSVD) and evaluated two different classifiers to verify our hypothesis. We selected a widely used dataset, which is described in the following.

2.3.1 ORL Database of Faces

We used the *ORL database* [57], which contains 400 images from 40 different subjects. These subjects are either students from Cambridge or Olivetti employees (in both cases male and female) and their age ranges from 18 to 81 with the majority between 20 and 35.

The database was collected from 1992 to 1994 and no restrictions were made regarding the facial expression. The recording conditions allowed only little side movement and limited tilt, hence only frontal faces are represented in the collection. Most of the subjects were photographed at different times with varying lightning conditions except the homogeneous dark background. Additional intra-class variation was caused by subjects with amblyopia. Some of the images shows them with and the rest of the subset without wearing their glasses. The images are encoded with 256 grey levels per pixel and each image has a 92×112 pixel resolution. Figure 2.2 shows an overview of all subjects in the ORL dataset. The dataset can be retrieved from the URL[2].

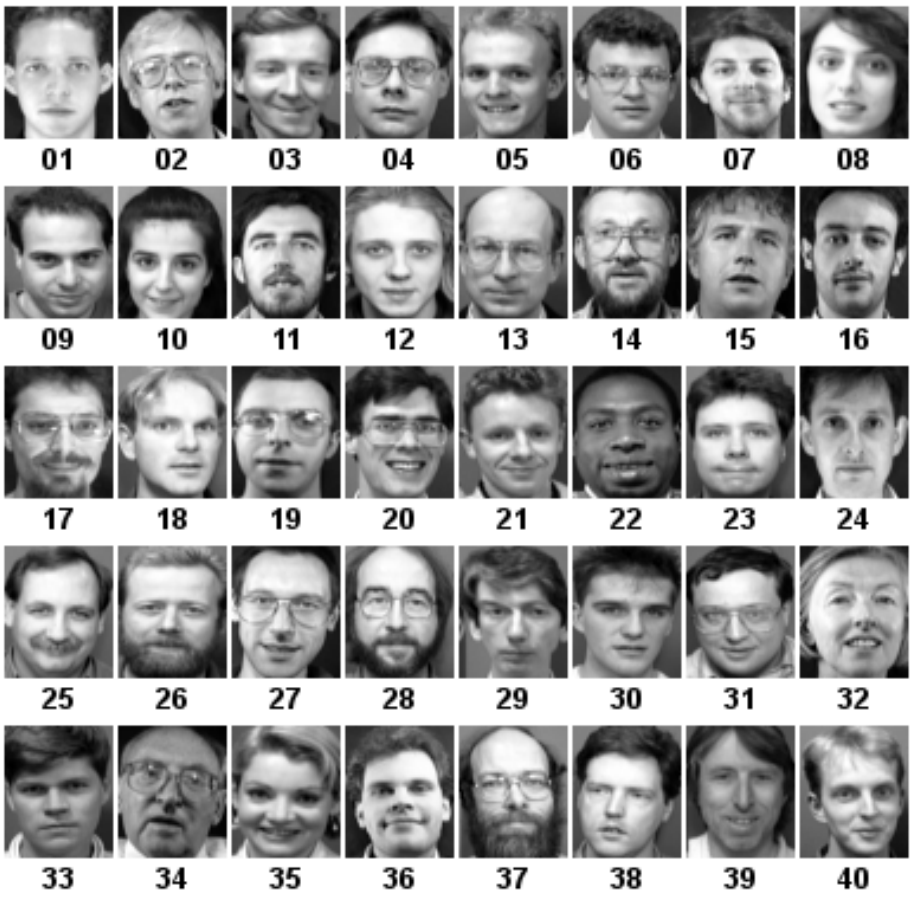


Figure 2.2: The ORL database of faces: It contains 400 images from 40 different subjects. The recording conditions were limited, thus only frontal faces are represented in the collection.

[2] `http://www.cl.cam.ac.uk/Research/DTG/attarchive:pub/data/att_faces.zip`

2.3.2 Sparse NMF using Structured Initialization

In the following we compared two methods to obtain a part-based representation. The multiplicative NMF algorithm [35] (eq. (2.2)) represents the baseline approach. We claim that Hoyer's sparse NMF extension [26] provides better representations with respect to the accuracy of the subsequent classification process. To provide a equitable comparison, we applied the NNDSVD approach to obtain a structured initialization of $\mathbf{W}$ and $\mathbf{H}$, for both of them.

Table 2.1 depicts the *natural* sparseness of the basis vectors $\mathbf{w}_i$, using the measurement defined in eq. (2.3). We reported the average and standard deviation over all basis $\mathbf{w}_i$: $i = 1, \ldots,$ number of modes. Obviously, when projecting the images to a low dimensional space, a small number of basis vectors have to encode the variance in the input data and the basis is seldom sparse. Matrix $\mathbf{W}$ becomes less compact, when the number of modes increase, but table 2.1 shows a slow change and the level of sparseness remains low.

modes	10	20	30	40	50	60	70	80
$\overline{\mathrm{sp}(\mathbf{w}_i)}$	$0.33 \pm .05$	$0.36 \pm .05$	$0.40 \pm .04$	$0.43 \pm .05$	$0.44 \pm .05$	$0.45 \pm .05$	$0.47 \pm .06$	$0.48 \pm .05$

Table 2.1: Average sparseness of the basis vectors $\mathbf{w}_i$ achieved using NMF with structured initialization of $\mathbf{W}$ and $\mathbf{H}$. An increased number of modes in $\mathbf{W}$ results in more sparse basis vectors. However, the effect is limited, since $0 \leq \mathrm{sp}(\mathbf{w}_i) \leq 1$.

Figure 2.3 illustrates 40 basis images using the NMF with structured initialization after 200 iterations. It shows facial parts like forehead, chin and eyes, but unlike expected these parts are not separated in one particular basis vector. This is evident with the observation summarized in table 2.1 — the sparseness of $\mathbf{W}$ is low.

We claimed that a structured initialization of $\mathbf{W}$ and $\mathbf{H}$ supports the consecutive sparse NMF towards a more part-based representation of the input data. In the following we compared the derived representations as well as the characteristics regarding convergence using random initialization and NNDSVD. The results will elucidate the advantages of a structured initialization, when used in combination with Hoyers's NMF method.

Figure 2.4 shows a comparison between basis images obtained from the sparse NMF approach. We fixed the number of modes to 40 and constrained the sparseness of the basis vectors $\mathbf{w}_i$ to a value of 0.8. The maximum number of iterations was set to 200. In case of a random initialization, the following sparse NMF algorithm derived the scattered basis vectors illustrated in fig. 2.4(a). The matrix factorization tends to optimize individual pixel locations rather than coherent regions. Figure 2.4(b) shows the basis images, when

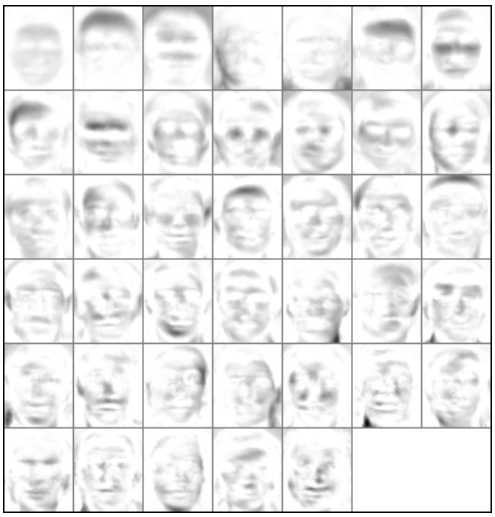
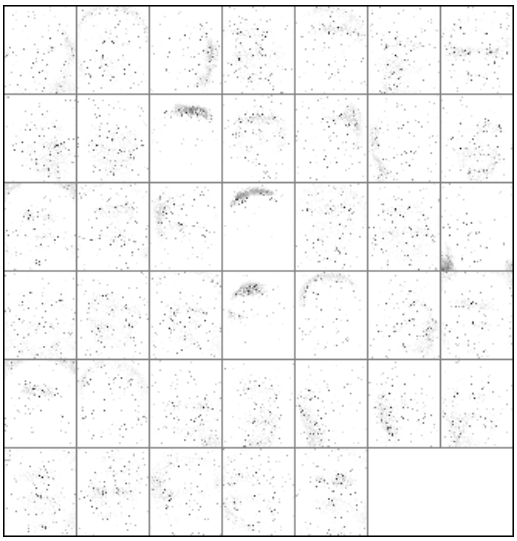
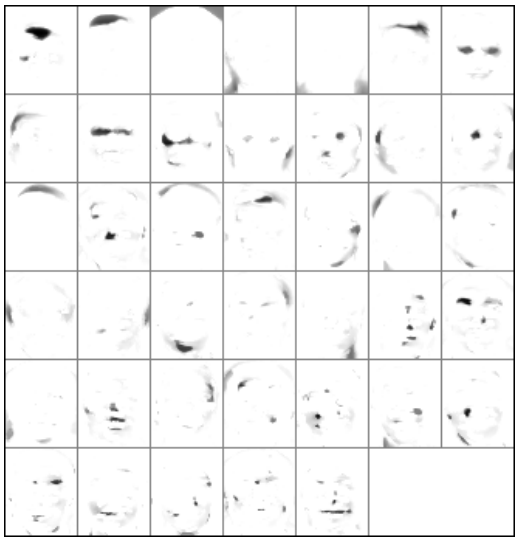


Figure 2.3: Basis images derived from the ORL database using NMF and structured initialization. The obtained representation is global, since multiple parts, like forehead, chin, eyes, etc., are present in each basis image.



(a) Basis images obtained using sparse NMF with random initialization of **W** and **H**.



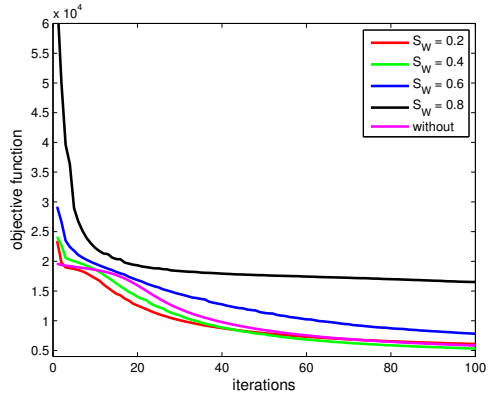
(b) Part-based representation derived from the sparse NMF approach using structured initialization (NNDSVD).

Figure 2.4: Comparison between basis images derived from the sparse NMF approach using 40 modes and 200 iterations. The sparseness constrain was set to 0.8. The structured initialization beneficially supports the NMF algorithm to obtain a better part-based representation.

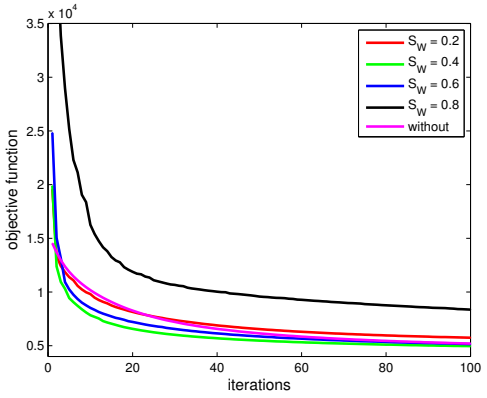
using the initial matrices $\mathbf{W}, \mathbf{H}$ obtained by a method proposed by Boutsidis and Gallopoulos [9] (NNDSVD). These results suggest, that a structured initialization supports the consecutive sparse NMF algorithm towards a much better part-based representation.

Figure 2.5 illustrates our last remark regarding initialization. The NNDSVD method provides starting conditions, that allows the subsequent NMF to significantly reduce the approximation error after few iterations. Each plot depicts the approximation error eq. (2.2) versus the number of iterations. Although, this property was already observed in [9], we investigated the issue using various levels of sparseness. Note that fig. 2.5(b) uses a different scaling on the Y-axis for better visualization. The line plot denoted as *without* (magenta) was generated using standard NMF.

Finally, we want to emphasize that a lower approximation error does not automatically guarantee better recognition results. Consider the objective functions on top of fig. 2.5(a) and fig. 2.5(b) (black line plot). Both of them were generated by limiting the sparseness of the basis vectors $\mathbf{w}_i$ to 0.8. In the next section we will illustrate that such a representation allows for better performance results, regardless of the considerable larger reconstruction error.



(a) Progress of the objective function using unconstrained and sparse NMF with random initialization.



(b) Progress of the objective function using unconstrained and sparse NMF with structured initialization (NNDSVD).

Figure 2.5: Different development of the approximation error (denoted as 'objective function') when comparing (left) random initialization vs. (right) structured initialization. Note the different scaling for the axis of the ordinate. Structured initialization leads to a faster reduction regardless of the sparseness level.

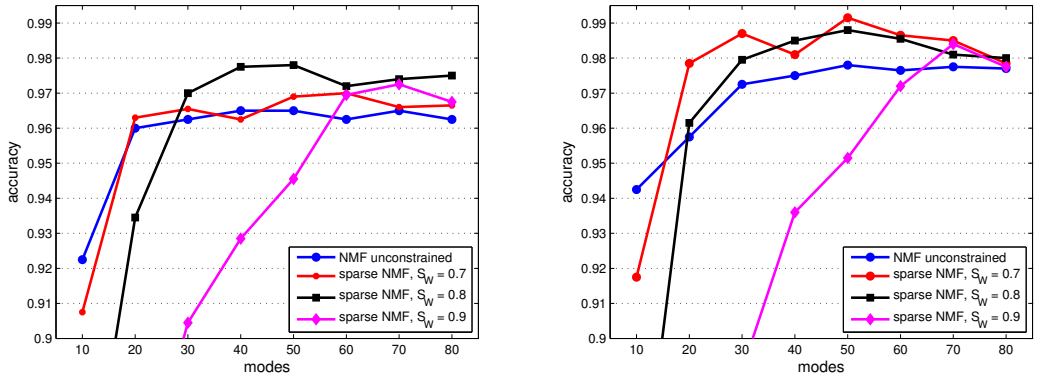
2.3.3 Recognition Results

We compared the recognition accuracy of a simple NN classifier using the Euclidean distance metric and the LMNN classifier described in section 2.2. Since the ORL dataset comprises ten different facial poses per subject, we performed a 10-fold cross validation. For each run, one pose per subject was left out for testing and all remaining images were used as training samples. During training we applied the NNDSVD method described in section 2.2 to obtain the initialization of $\mathbf{W}$ and $\mathbf{H}$. Afterwards, we performed the NMF as well as the sparse NMF approach on the trainings data. Due to the deterministic nature of the initialization process, multiple runs of the matrix factorization could be avoided. We performed multiple runs using different number of basis vectors $\mathbf{w}_i$, to investigate the influence of the feature dimension. Furthermore, we evaluated different constraints on $\mathbf{W}$ using the sparse NMF method, to verify our hypothesis. Once a representation is found, we projected the test images using the same iteration rules already explained in section 2.2, while keeping the matrix of basis vectors $\mathbf{W}$ constant. The task of the final classifier was the correct identification of the subject from an unseen pose image.

Figure 2.6(a) depicts the recognition performance for a NN classifier. In case of an unconstrained NMF representation (blue curve) no significant differences in the classification performance could be recognized, if the number of basis vectors exceeds 20. However, when using sparse features and a sufficient number of modes, the accuracy could be further improved. For example, using 50 base dimensions and setting $S_W = 0.8$ increased the accuracy by $\approx 1.25\%$. This emirically demonstrates that a sparse representation beneficially supports the following classification process.

However, fig. 2.6(a) also depicts the correlation between sparseness and feature dimensionality. Too few basis vectors and a strong constraint regarding sparseness led to a representation, which is not distinctive enough. This effect is illustrated for $S_W = 0.9$ and modes < 60 as well as for $S_W = 0.8$ and modes < 30.

We observed a similar behaviour using the LMNN classifier. During training $k = 3$ nearest neighbours were used to learn the metric, as suggested in [69]. Figure 2.6(b) summarizes the achieved recognition performance using the same evaluation schema as in fig. 2.6(a). Again, a classifier using the sparse representation is able to perform better, if the features space is large enough. Figure 2.6(b) also indicates that the optimal dimension of the feature space lies between 30 and 60, when using a sparseness level of 0.7 or 0.8. For feature dimensions ≥ 70 no significant improvement could be observed, compared to the unconstrained representation. We hypothesize that increasing the number of modes might



(a) Recognition performance of a nearest neighbour classifier using unconstrained and sparse NMF representations.

(b) The accuracy of the LMNN classifier ($k = 3$) could be improved using sparse features ($S_W = 0.7 - 0.8$).

Figure 2.6: A sparse representation is able to support the subsequent classifier. A NN classifier utilizing the Euclidean distance metric (left) and the LMNN classifier proposed by [69] (right) performed better, when using features obtained by NMF with sparseness constraints. The baseline in both plots (blue curve) constitutes an unconstrained NMF.

decompose meaningful parts and the obtained encoding is therefore more ambiguous for the consecutive classifier.

2.4 Conclusions

In this chapter, we demonstrated that a sparse part-based representation increases the performance of the subsequent classifier. First, we showed that the features obtained by using Hoyer's NMF with sparseness constraints [26] are more compact in contrast to the multiplicative NMF algorithm proposed by Lee and Seung [35]. Second, when using a structured initialization method (NNDSVD [9]) instead of the widely used random initialization, the subsequent sparse NMF derives a better part-based encoding of the input data.

In the context of HAR, such a part-based representation might be beneficial. The proposed combination of structured initialization and sparse NMF decomposes the entire feature vector, representing the human body, into multiple compact parts (e.g., limbs, head). Figure 2.4(b) illustrates this effect in the scope of faces, where black regions denote non-zero elements in the basis vector. One particular part is represented by multiple

attributes in $\mathbf{w}$, contrary to fig. 2.4(a), where only few attributes encode this region in the underlying image.

Finally, we confirmed our initial hypothesis by comparing the recognition accuracy using differently sparse representations. In particular, we evaluated two different NN classifiers (NN vs. LMNN [69]). For both classifiers we observed the following behaviour. If the level of sparseness S_W of the basis vectors in $\mathbf{W}$ is significantly greater than those obtained by unconstrained NMF, then a better recognition accuracy could be achieved. Nevertheless, the results also exhibited the limitations of sparse features. If the dimensionality of the feature space was low, and the sparseness constraint too strong, then we observed a significant reduction of the classification performance. We hypothesize that the obtained representation was not sufficiently rich to encode the variations of the original data. Since the number of basis vectors $\mathbf{w}_i$ was low and they had to be sparse, this consequently led to a reduced number of parts in the final encoding. For example imagine an extreme condition having only three modes, e.g., the nose, the eyes, and the chin. It seems unlikely that a classifier successfully discriminates a larger number of individual faces.

I hear and I forget. I see and I remember.
I do and I understand.

Confucius

Chapter 3

Multiple Instance Learning for Human Action Recognition

Contents

3.1 Multiple Instance Learning 24
3.2 MIL Algorithm Reviewed 27
3.3 Comparative Study 36
3.4 Experimental Results 44

We have already motivated MIL in the scope of action recognition (see section 1.1). By the use of this concept, the effort for preparing suitable trainings data could be reduced. The labelling of individual samples is relaxed to labelling of bags (sets comprising multiple samples). Furthermore, when using object detection processes, extensive filtering of false positives or selection of adequate detection results can be omitted.

MIL has drawn much attention in the computer vision community, and different fields of applications have been identified. Viola et al. [64] for instance proposed a multiple instance boosting algorithm (MILBoost) and evaluated it in an object detection task. Within their application, they treated images as bags and a sub-region depicting a persons head (the object) as positive instance. Sub-windows, which did not overlap with the object, were considered negative. Viola et al. demonstrated superior performance results compared to a AdaBoost framework [63].

Babenko et al. [6] applied MIL in the context of visual tracking. The adaptive appearance model of the tracker is updated using the concepts of bags and instances. The ambiguity in the data originates from the updated tracker position. In particular, positive

samples are extracted within an ϵ-environment of the current tracking location. Negative ones are cropped out from distant regions in the actual frame. The proposed method outperformed other online tracker methods and achieved real-time performance.

Roth et al. [54] introduced a robust multiple camera learning approach based on MIL. They focused on co-training of several classifiers and utilized the homography information to exchange detection results from distinct camera views. Their efficient and robust concept collaborates only during training and performance results where presented for the task of person detection. They showed the generality of the method on different standard datasets with varying complexity and achieved superior results compared to state-of-the-art detector approaches.

Differently to the before mentioned approaches, uncertainty in the context of action recognition consists of two distinct parts. First, the location of the action in one single frame is unknown, i.e., where in the image is the person performing the action. Second, the temporal position of the action in the sequence is unknown. One can only assume that the action is present. The ambiguity is passed on to the MIL algorithm, which now has to figure out an appropriate decision boundary. The promising results of the previous approaches in computer vision motivated us to apply MIL also in the context of HAR.

In the following we introduce the mathematical formulation of MIL in section 3.1 and discuss the differences to standard supervised learning. One major objective of this thesis is to compare different MIL approaches against each other. Therefore, we discuss five selected methods in section 3.2 and present experimental results in section 3.4. A detailed comparison of diverse approaches in the context of HAR concludes each individual benchmark.

3.1 Multiple Instance Learning

The term *Multiple Instance Learning* (MIL) was originally introduced by Dietterich et al. in the context of drug discovery [14]. Several years before Dietterich, in the early 1990 Keeler et al. [29] originated the spirit of Multiple Instance Learning. Starting with the 'axis-parallel rectangle' algorithm in [14], the MIL paradigm has emerged as learning algorithm. One significant contribution in this field was introduced by Maron et al. [41]. Their Diverse Density (DD) algorithm was claimed as a general framework for solving the MIL problem. Andrews et al. [3] modified the standard *Support Vector Machine* (SVM) approach for the MIL paradigm and formalized a statistical learning framework that is able to handle ambiguous examples.

The next section emphasizes the differences between MIL and supervised learning and introduces the mathematical formulation of this alternative learning concept.

3.1.1 Supervised vs. Multiple Instance Learning

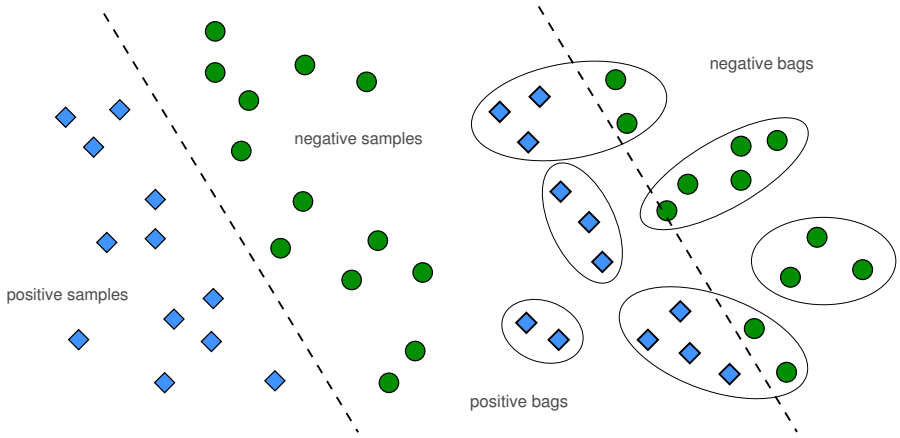
As mentioned at the beginning of this chapter MIL can be seen as one alternative to the supervised learning concept, particularly, when dealing with ambiguous labels. The main difference lies in the way how labels are handled during training. To clarify this, we will review the supervised learning scenario and discuss its advantages and drawbacks in comparison with MIL.

Supervised Learning: In supervised learning each example is an instance augmented with a class label. Given a set of examples and their corresponding labels, the task of the algorithm is to learn a function that predicts the label for unknown examples. The output of the algorithm is a classifier or model of the data. Assuming binary class labels the supervised learning scenario can be formally described as follows:

$$\begin{array}{lll} \mathcal{X} = \{x_1, \ldots, x_n\} & x_i \in \mathbb{R}^d & \textit{example} \\ \mathcal{Y} = \{y_1, \ldots, y_n\} & y_i \in \{0, 1\} & \textit{label} \\ h(\mathcal{X}) : \mathcal{X} \mapsto \mathcal{Y} & & \textit{classifier} \end{array} \tag{3.1}$$

Figure 3.1(a) depicts a simple 2D toy example where all positive instances (blue diamond) are separated from the negative samples (green circle) through the dashed line illustrating the decision boundary of the classifier $h(\mathcal{X})$. It is indispensable that all samples $x_i \in \mathbb{R}^d$ have a corresponding label $y_i \in \{0, 1\}$ assigned during training. Therefore, thoroughly labelled training data is the prerequisite for supervised learning.

Multiple Instance Learning: The following simple MIL problem adapted from [14] explains the key concept of Multiple Instance Learning and is illustrated in Fig. 3.2. Suppose there are several faculty members, and each of them owns a key chain that comprises a few keys. Some members are able to enter a restricted area (these members possess a 'positive' key chain) and some are not (virtually 'negative' labelled key chain). The exercise is to predict whether the investigated key chain gains access to the secret room. The solution is to find the key that all 'positive' key chains have in common. Being able to correctly identify the key that gives access to the restricted area, any other key chain can be classified correctly – either it contains the required key, or not.



(a) Supervised Learning (b) Multiple Instance Learning

Figure 3.1: Different learning paradigms shown for a two dimensional two-class toy problem. The dashed line illustrates the decision boundary of the classifier for (a) supervised learning and (b) multiple instance learning; each ellipse depicting a bag.

Figure 3.2: Bunch of keys illustration as Multiple Instance Learning (MIL) example (adopted from [14], reprinted from [4])

3.1.2 Formalization of the MIL concept

MIL introduced the concept of bags and the a dataset $\mathcal{X}$ is organized in bags. Each bag $\mathbf{B}_i$ consists of an arbitrary number of samples $x_{ij} \in \mathbb{R}^d$. These samples are also called instances. Since the MIL problem was originally defined for binary cases, suppose a two-class problem with labels $y_i \in \{0, 1\}$. This leads to

$$\begin{aligned} &\mathcal{X} = \{\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_n\}; \quad \mathbf{B}_i = \{x_{i1}, \ldots, x_{im}\}; \quad x_{ij} \in \mathbb{R}^d \quad \textit{bags} \\ &\mathcal{Y} = \{y_1, \ldots, y_n\}; \qquad\qquad\qquad\qquad\quad y_i \in \{0, 1\} \quad \textit{label} \end{aligned} \tag{3.2}$$

The label of bag $\mathbf{B}_i$ is derived via the standard MIL assumption (see eq. (3.3)). Negative bags consist only of negatively labelled instances, while bag $\mathbf{B}_k$ is labelled positive if it contains at least one positive instance $x_{kj} : y_{kj} = 1$. These constraints can be formalized as

$$\begin{aligned} \mathbf{B}_k \quad &\text{is pos. iff} \quad \exists\, x_{kj} \quad \text{with } y_{kj} = 1, \quad x_{kj} \in \mathbf{B}_k \\ \mathbf{B}_k \quad &\text{is neg. iff} \quad \underset{j}{\forall}\, x_{kj} \quad \text{with } y_{kj} = 0, \quad x_{kj} \in \mathbf{B}_k \end{aligned} \tag{3.3}$$

and are known as standard MIL assumption. However, there exist more generalized instance-based assumptions for MIL, e.g., presence-based (related to the key chain example (see fig. 3.2) which would imply doors with several locks), threshold-based (where a bag $\mathbf{B}_k$ is labelled 'positive' if at least c instances have a positive label), or a count-based MIL assumption. Weidmann et al. formed a concept hierarchy of these assumptions in [68]. Different to the supervised learning scheme in eq. (3.1) the input of the learning algorithm is a set of labelled bags. The labels of individual instances are not known during training. It is even more ambiguous, due to the fact that within positive bags no information about negative, or even unlabelled instances is present. Equation (3.2) and eq. (3.3) summarize the problem statement formally. The task of every MIL algorithm is to train either an instance classifier $h(x_{ij})$ or a bag classifier $H(\mathbf{B}_i)$ and formally denoted as

$$h(x_{ij}) : \mathcal{X} \mapsto \mathcal{Y} \quad \textit{instance classifier} \tag{3.4}$$

$$H(\mathbf{B}_i) : \mathcal{X}^m \mapsto \mathcal{Y} \quad \textit{bag classifier}. \tag{3.5}$$

However it depends on the algorithm which kind of classifier is learned, but the majority of them formulate the instance classifier in eq. (3.4). Anyway, it is straightforward to construct a bag classifier from an instance classifier via $H(\mathbf{B}_i) = \max_j h(x_{ij})$.

Table 3.1 summarizes the notation used in this work regarding MIL and supervised learning. We adopted the formalization basically from [4] and follow it, except when otherwise specified, during the algorithmic description of different approaches.

3.2 MIL Algorithm Reviewed

In this section we briefly review the different MIL approaches used in our comparative study. We begin with the formal description of each algorithm, continue with their characteristic properties and finish each section with a discussion about their advantages and drawbacks. In particular, we consider the question *why* this specific approach is worth

$x_i \in \mathbb{R}^d$	i^{th} instance (supervised)
$\mathbf{B}_i : \{x_{i1}, x_{i2}, \ldots, x_{im}\}$	i^{th} bag with m instances x_{ij} (MIL)
$y_j \in \{0, 1\}$	label of i^{th} instance (supervised) or i^{th} bag (MIL)
y_{ij}	true label of j^{th} instance in i^{th} bag
n	number of instances (supervised) or bags (MIL)
m	number of instances per bag
d	dimensionality of the feature space
$h(\mathcal{X})$	supervised classifier
$h(x)$	instance classifier (MIL)
$H(\mathbf{B})$	bag classifier (MIL)

Table 3.1: A summary of notation used to describe the basic concept of multiple instance learning and the algorithms used in this work.

being selected in this work.

3.2.1 Kinematic-MIL

Ali and Shah [2] proposed a new kind of kinematic features set and showed improved performance by using MIL in an human action recognition task. The contribution of their work is twofold. First they developed a new set of features that are derived from the optical flow of a sequence of images. They call them *kinematic* features motivated by the assumption that different aspects of dynamic motion patterns within the optical flow field can be represented more discriminative compared to the standard flow. As our work does not focus on representation, we refer the reader to their work (see [2] sec. 3) for a detailed description.

The second and more important contribution of Ali and Shah related to our work is their computation of *kinematic modes* and the application of multiple instance-based learning in the context of human action recognition. Emanating from kinematic features they extracted kinematic modes from every single video. This is achieved performing *Principal Component Analysis* (PCA) on the spatio-temporal volumes of each feature channel separately. They claimed that such a representation is sufficiently rich to encode the essential patterns of the human action, because each video presents only one action. They claimed that such an representation is sufficiently rich to encode the essential patterns of the human action, because each video presents only one action.

According to the notation in table 3.1 let

$$\mathbf{B}_i = \{x_{i1}^{f^1}, \ldots, x_{in}^{f^1}, x_{i1}^{f^2}, \ldots, x_{in}^{f^2}, \ldots, x_{i1}^{f^k}, \ldots, x_{in}^{f^k}\}$$

denote the i^{th} bag, $x_{ij}^{f^k}$ represents the j^{th} kinematic mode in $\mathbf{B}_i$ (also known as instance see the general formulation in section 3.1.2), and the superscript f^k indicates the type of kinematic mode with $k = (1, \ldots, 11)$. Following the spirit of the diverse density framework originated by [14] and [12] the authors propose a *kinematic-mode based* embedding in a high dimensional space $\mathbb{F}_C$, where C denotes the set of all kinematic modes $x_{ij}^{f^k}$ from all trainings bags regardless of their label. Note that all modes in C are re-enumerated without consideration of their bag-origin as $x_1^{f^1}, \ldots, x_e^{f^{11}}$. The conditional probability of an attribute in C belonging to one bag is then given by

$$P(x_e^{f^k}|B_i) \propto d(x_e^{f^k}, B_i) = \max_j exp\Big(- \frac{||x_{ij}^{f^j} - x_e^{f^k}||^2}{\sigma^2}\Big), \forall f^j = f^k. \tag{3.6}$$

Within this single instance space $\mathbb{F}_C$ each bag is represented by exactly one point $m(\mathbf{B}_i)$ and is formally written as

$$m(\mathbf{B}_i) = \Big[d(x_1^{f^1}, \mathbf{B}_i), d(x_2^{f^1}, \mathbf{B}_i), \ldots, d(x_e^{f^{11}}, \mathbf{B}_i)\Big]^T. \tag{3.7}$$

A simple nearest neighbour classifier in the single instance space $\mathbb{F}_C$ is learned to predict labels for the test bags.

Discussion: One interesting novelty of this approach is the use of PCA to encode spatio-temporal patterns within each *single* video. We suspect that this additional step in encoding individual action creates a kind of mid-level representation that allows more robust classification. One the other hand, it is questionable, if such a representation is sufficient for more complex actions, since, the number of modes for each feature type is constant for all actions [2]. It seems obvious that a simple 'single hand wave' action can be encoded more compactly as an asymmetrical 'serve' at tennis. Another possible disadvantage is that the embedding depends on the number of bags used during training. Remember that each bag $m(\mathbf{B}_i)$ is embedded in $\mathbb{F}_C$. The dimensionality of $\mathbb{F}_C$ is $|C|$ that is 'number of modes per feature' $\times$ 11 $\times$ 'number of bags'. If a large number of training-bags is used for such an classifier, the embedding step is slow and requires excessive memory in both training- and testing-phase. Finally, we selected this approach, because it is the only one that originally evaluated its potential on benchmark datasets for human action recognition.

3.2.2 miGraph

Motivated by the assumption that instances in a bag are not *independently and identically distributed* Zhou et al. [77] proposed two alternative MIL approaches. Both take advantage of the observation that many real objects are a structured composition of several parts. Their framework is able to determine the inherent configuration between instances representing such objects. This property supports the classification process, thus making it more reliable and accurate.

Figure 3.3 depicts this idea of bags with instances not independently and identically distributed. Each diamond is one instance in the bag. Note that during training no knowledge concerning the true labels of instances is available. Thus, we illustrated all samples omitting the label information. A solid line represents a stronger relation among instances compared to the dashed line. Similar principles have been proposed in the scope of object recognition and are known as constellation models, i.e. [67]. In order to avoid an explicit model of parts, the relations in Zhou et al. [77] are learned during training. Assume the left bag and the bag in the middle of fig. 3.3 being more similar than the bag on the right side. The number of instances in the bags $\mathbf{B}_i$ are equal and the position of the instances x_{ij}, representing the location in $\mathbb{R}^2$, are similar in all three bags. It is obvious to see that ignoring the inherent configuration of the instances in each bag makes the classification task difficult.

Zhou et al. evaluated both approaches on text and image categorization as well as multi-instance regression. Their experiments support the hypothesis that relations among instances in a bag conveys important information. Furthermore, they demonstrated that the second method called *miGraph* performed better and is less computational complex in contrast to the other method called *MIGraph*, especially when the size of the bags is large. Thus, in our work we focus on *miGraph* and give a more formal description of the algorithm in the following paragraph.

For *miGraph* an affinity matrix W^i is derived for every bag $\mathbf{B}_i$. This is done by comparing the pairwise distances between all instances x_{in} and x_{im} in the bag with a predefined threshold δ. The element w_{nm} is set to 1 if this distance is smaller than δ, and 0 otherwise. The authors used Gaussian distances in their algorithm and set δ^i to the average distance in the bag $\mathbf{B}_i$:

$$w_{nm}^i = \begin{cases} 1 & \text{iff } exp(\frac{||x_{in}-x_{im}||^2}{2\sigma^2}) \leq \delta^i \\ 0 & \text{otherwise.} \end{cases} \tag{3.8}$$

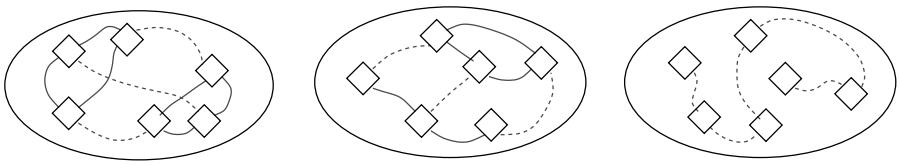


Figure 3.3: Illustrating that relations among instances within bags can help classifying them. The left bag and the bag in the middle are more similar than the bag to the right. Note that a solid line between instances indicate a stronger relation compared to the dashed line (reproduced from [77]).

Given the affinity matrix for two bags $\mathbf{B}_i$ and $\mathbf{B}_j$ in *miGraph* a kernel $k_g(\mathbf{B}_i, \mathbf{B}_j)$ is defined to express the similarity between the bags. This kernel is formulated for $\mathbf{B}_i$ which contains n_i instances and $\mathbf{B}_j$ that consists of n_j instances. To clarify, w^i_{au} is the element in W^i at the a^{th} row and u^{th} column. Using this notation the kernel is defined as

$$k_g(\mathbf{B}_i, \mathbf{B}_j) = \frac{\sum_{a=1}^{n_i} \sum_{b=1}^{n_j} W_{ia} W_{jb}}{\sum_{a=1}^{n_i} W_{ia} \sum_{b=1}^{n_j} W_{jb}} \, k(x_{ia}, x_{jb}), \tag{3.9}$$

where $W_{ia} = 1/\sum_{u=1}^{n_i} w^i_{au}$, $W_{jb} = 1/\sum_{v=1}^{n_j} w^j_{bv}$, and $k(x_{ia}, x_{jb}) = exp(\frac{||x_{ia}-x_{jb}||^2}{2\sigma^2})$ is defined similar to the distance metric in eq. (3.8). In order to solve the classification problem, this kernel is fed into support vector machines.

Discussion: We strongly suspect that every human action has its own inherent structure, especially more complex ones, where i.e. multiple limbs are included. The results proposed by Zhou et al. [77] provide a strong indication that utilizing the relations among instances supports the final classification process. When applying such an approach to human action recognition, we assume that these correlations can be represented in W^i and as a consequence making the classification less ambiguous. One disadvantage of the method is its computational complexity, i.e., obtaining the kernel k_g requires $O(n_i \, n_j)$.

3.2.3 MILES

Another MIL approach that is based on the diverse density framework [40] is called *Multiple Instance Learning via Embedded Instance Selection* (MILES) and was proposed by Chen et al. [12]. The embedding is based on the instances in every bag, but in contrast to Maron's work [39] they do not assume the existence of one single target concept. The MIL assumption used in MILES is more 'general' [19] because bag labels are derived by

the relationship of instances to a set of target points. This is contrary to the count-based assumption used in the Citation k-NN approach (see section 3.2.5), where R references and C citations determine the final bag label. Section 3.1.1 summarizes the discussion on various MIL assumptions.

In the following we briefly review the basic concepts of MILES. Similar to the diverse density framework, bags $\mathbf{B}_i$ are embedded into a single instance feature space $\mathbb{F}_C$ by estimating the probability that one instance is a target concept x^k,

$$Pr(x^k|B_i) \propto s(x^k, B_i) = \max_j exp\Big(- \frac{\|x_{ij} - x^k\|^2}{\sigma^2}\Big), \tag{3.10}$$

regardless of the bags's class label, where x_{ij} are instances in bag B_i, σ is a predefined scaling factor. The term $s(x^k, B_i)$ describes the similarity between a single target concept x^k and the bag $\mathbf{B}_i$. Chen et al. defined $\mathbb{C} = \{x^k : k = 1, \ldots, n\}$ to be the concept class and denoted x^k as one target concept out of $\mathbb{C}$. In other words, $\mathbb{C}$ is simply a set of re-indexed instances from all positive and negative bags that are used during training. This means that no artificial target point is calculated, as the centroid of each individual bag, but MILES assumes that the target concepts can be approximated by all instances presented during training. Note that the size of $\mathbb{C}$ defines the dimensionality of $\mathbb{F}_C$, which can be high for real world problems with many bags and/or many instances per bags. This motivated the authors to simultaneously perform feature selection and classification by the use of one-norm SVMs.

The concept of one-norm SVMs is to reformulate the standard quadratic optimization problem to a linear one [76, 78, 79]. Such a linear program can be solved efficiently, even in high dimensional feature spaces. One alternative formulation [78] utilizes the 'Manhattan length' (also known as 'City block distance') of the weight vector $\mathbf{w}$ to define the penalty term as $\|\mathbf{w}\|_1 = \sum_k |w_k|$. The additional benefit of the one-norm SVM is that most of the features are set to zero. Chen et al. used this property to simultaneously determine a sparse representation and construct a fast classifier for bags. For completeness we summarize the formulation of the resulting instance classifier in eq. (3.11) and refer the interested reader to [12] for a comprehensive discussion on MILES. Assume $\mathcal{I} = \{k : |w_k^*| > 0\}$ as the index set for non-zero entries in the optimal solution $y = \text{sign}(\mathbf{w}^{*T}\mathbf{m} + b)$ with $\mathbf{m} \in \mathbb{R}^{|C|}$ representing a bag in features space $\mathbb{F}_C$ (see eq. (3.7)). The instance classifier is

given as

$$h(\mathbf{B}_i) = \operatorname{sign}\left(\sum_{j^* \in \mathcal{U}} g(x_{ij^*}) + b\right), \quad \text{with} \tag{3.11}$$
$$g(x_{ij^*}) = \sum_{k \in \mathcal{I}_{j^*}} \frac{w_k^* \, s(x^k, x_{ij^*})}{m_k}.$$

The index set $\mathcal{U}$ defines a minimal set of instances responsible for the classification of bag $\mathbf{B}_i$. Since one-norm SVMs favours sparse representations only a few features $s(x^k, x_{ij^*})$ have to be evaluated. This makes the labelling of unseen instances efficient from a computational perspective.

Discussion: Chen et al. demonstrated good performance and high robustness to label noise in the scope of region based image classification as well as object class recognition experiments [12]. The major advantage of MILES compared to the other methods selected in this paper is that once trained, classification of test bags can be performed very fast. By evaluating only very few features $|w_k^*| > 0$ from instances in the test bag, this method is effective and computational efficient. Another reason to include this approach in our study is its ability to predict both bag and instance labels. This could be a decisive benefit in some applications. A more diversified viewpoint of this approach was given by Foulds and Frank [20] who classified MILES as a general wrapper algorithm to emphasize that the original one-norm SVM could easily exchanged by other learning techniques like C4.5 [52], random forests [11] or standard SVMs. Finally, it is notable that MILES inspired later MIL approaches e.g. [2] (see section 3.2.1).

3.2.4 MI-SVM

Andrews et al. [3] proposed two approaches to modify *Support Vector Machines* (SVM) for MIL problems. Compared to other methods like MI-kernel based algorithms (see section 3.2.2) they hypothesize that a more conceptual modification of SVMs is necessary to deal with MIL settings. The authors evaluated their approach on different problem statements like automatic image annotation or text categorization. Based on that we favour *MI-SVM*, which is suitable for problems where only bag labels are concerned. It is therefore sufficient for a human action recognition task, since we assume that videos are represented as bags and the goal is do determine whether a specific action is present or not.

The key idea of *MI-SVM* is an alternative *maximum bag margin formulation* shown in eq. (3.12). Here, for positive bags $\mathbf{B}_i : H(\mathbf{B}_i) = 1$ a selector variable $s(\mathbf{B}_i) \in \mathbf{B}_i$ denotes a single instance x_{ij} as the 'witness' in $\mathbf{B}_i$. This single instance determines the margin of the bag. Identifying this instance becomes critical, since the resulting mixed integer programming problem cannot be solved efficiently. Andrews et al. formulated a heuristic strategy to determine these selector variables $s(\mathbf{B}_i)$ for positive bags:

$$\min_{s} \min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i \tag{3.12}$$

$$\begin{aligned} s.t. \quad & \forall\, \mathbf{B}_i : H(\mathbf{B}_i) = -1 \quad \text{and} \quad -\langle \mathbf{w}, x_{ij}\rangle - b \geq 1 - \xi_i, \quad \forall x_{ij} \in \mathbf{B}_i, \\ & \forall\, \mathbf{B}_i : H(\mathbf{B}_i) = +1 \quad \text{and} \quad \langle \mathbf{w}, x_{i\,s(\mathbf{B}_i)}\rangle + b \geq 1 - \xi_i, \\ & \text{with} \quad \xi_i \geq 0. \end{aligned}$$

The algorithm is initialized by the centroid of the bag. During each iteration the approach tries to find the optimal discriminant function and recalculates the selector variable $s(\mathbf{B}_i) = \arg\max_{j \in \mathbf{B}_i} \langle \mathbf{w}, x_{ij}\rangle + b$ for every positive bag. This procedure stops, if the selector variables remain unchanged.

Discussion: The iterative nature of this algorithm is one possible drawback, especially, when addressing time critical problems. The heuristic strategy is another one, since it is likely to get stuck in a local minima. Our decision for MI-SVM over mi-SVM is well-founded in the arguments from Andrews et al.. They prefer MI-SVM when classifying new bags is beneficial compared to a more accurate instance classifier. Such circumstances apply in the scope of human action recognition, when we want to evaluate whether an action is present in a particular part of a video or not. The correct classification of individual frames is no primary goal in this context.

3.2.5 Citation k-NN

Wang and Zucker [66] proposed a modified version of the popular *k-nearest neighbour* (k-NN) approach in the scope of multiple instance learning. Their method is based on a framework used in the field of library search and information science. They adapted the notation of *reference* and *citation* to define the relevance between bags. In addition to derive the label of bag $\mathbf{B}_i$ from its neighbouring bags, bags that identify $\mathbf{B}_i$ as their neighbour are considered as well. The authors showed competitive but more robust results compared to [40] or [14] on standard benchmark datasets.

Clearly any nearest neighbour algorithm heavily depends to the definition of a distance metric. In [66] a modified *Hausdorff distance* is utilized to measure adjacencies between bags. Although, the Hausdorff distance provides such a metric function between subsets, in the case of MIL two sets of instances, it is known to be sensitive to single outlying points. Wang and Zucker formulated a variation of the standard metric and called it the *minimal Hausdorff distance* (see eq. (3.13)), that is empirically proven to be more robust with respect to noise:

$$H(\mathbf{B}_i, \mathbf{B}_j) = h_1(\mathbf{B}_i, \mathbf{B}_j) = h_1(\mathbf{B}_j, \mathbf{B}_i) \quad \text{with} \tag{3.13}$$
$$h_1(\mathbf{B}_i, \mathbf{B}_j) = \min_{a \in \mathbf{B}_i} \min_{b \in \mathbf{B}_j} ||x_{ia} - x_{jb}||.$$

Figure 3.4 illustrates the concept of reference and citation in the *Citation k-NN* approach using a simple 2D example. Suppose there are five bags $(\mathbf{B}_2, \ldots, \mathbf{B}_6)$ augmented with either a positive or a negative label that derives from the instances x_{ij} in the bags. Consistent with the previous notation (see fig. 3.1), a green circle denotes a negative instance, while a blue diamond denotes a positive one. Consider the task of predicting the label of an unseen bag $\mathbf{B}_1$ that contains several instances (illustrated as white triangles to indicate that their true label is not known). Without loss of generality, we investigate the R^{th} - nearest references and the C^{th} - nearest citers of $\mathbf{B}_1$ with $R = 2, C = 2$. Bag $\mathbf{B}_2$ and $\mathbf{B}_3$ are the two nearest neighbours of $\mathbf{B}_1$ and therefore called references. The outward-pointing arrows depict those relations in the figure. In the concept of citers the viewpoint changes from $\mathbf{B}_1$ to bags that call $\mathbf{B}_1$ as their nearest neighbours. Thus, in 3.4 the citers of $\mathbf{B}_1$ are $\mathbf{B}_2$, $\mathbf{B}_3$, and $\mathbf{B}_4$. This relation is depicted as inward-pointing arrow. The final decision on the bag label is based on the total number of positive bags p and negative bags n in **both** sets references and citations. If $p > n$ the bag $\mathbf{B}_1$ is classified positive, otherwise negative. In our example the R-nearest references of bag $\mathbf{B}_1$ are $\{\mathbf{B}_2, \mathbf{B}_3\}$, and the C-nearest citers are $\{\mathbf{B}_2, \mathbf{B}_3, \mathbf{B}_5\}$. As a consequence, bag $\mathbf{B}_1$ is labelled negative because $p = 2$ and $n = 3$.

Furthermore, fig. 3.4 points out the difference to standard k-NN majority voting. Consider the k nearest neighbours of bag $\mathbf{B}_1$ assuming $k = 3$. The three red solid arrows illustrate this situation. Following the considerations of k-NN voting the majority of the neighbouring bags are positive, contrary to our previous observation. Note regarding the underlying distribution of instances the bag $\mathbf{B}_1$ is more likely to be negative.

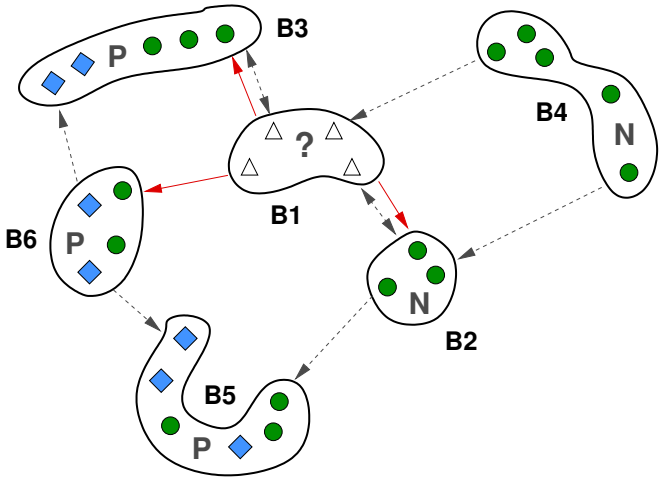


Figure 3.4: Citation k-NN approach compared to standard k-NN method. The example was inspired by Wang and Zucker [66].

Discussion: Wang and Zucker proposed their algorithm as highly competitive and alternative learning concept in the domain of multiple instance learning. One disadvantage might be the fact that in contrast to other algorithms like MILES (see section 3.2.3) or miGraph (see section 3.2.2) Citation k-NN is not able to predict the labels of individual instances. Although the method is known to be inefficient, we included this method in our comparative study because of its simplicity and the relation to the well known standard k-NN algorithm.

3.3 Comparative Study

One objective of this work is a comprehensive analysis whether the MIL concept is suitable for the action recognition task. For that reason we evaluated the previously presented MIL methods on two standard benchmark datasets. Each individual algorithm deals with exactly the same feature set. Furthermore, we focus on a consistent evaluation schema that guarantees comparable results. These are shown in section 3.4 and an extensive comparison with previously published results is given at the end.

HAR can be performed on different levels of abstraction [46] and clear restrictions are necessary to narrow the field of possible solutions. In our work we assume full-body movements and do not examine problems concerning partial occlusions. Furthermore, we concentrate on single body movements and do not deal with interactions between multiple human beings [56] or single humans and objects.

Note, within this chapter we focused on the evaluation of different *MIL classifiers* in the context of HAR. This is clearly in contrast to Wang et al. [65], who investigated various local *spatio-temporal features* in the scope of action recognition.

3.3.1 Histogram based feature representation

The combination of appearance and motion cues in the context of human action recognition has been suggested in several publications. Motivated by recent findings in cognitive science Jhuang et al. [27] and Schindler and van Gool [58] proposed independently that shape and motion cues, similar to the ventral and dorsal pathway in the visual cortex, complement each other and thus provide a richer representation to allow for robust classification of action. Authors like Laptev et al. [33] extracted sparse space-time features [32] in a multi-scale pyramid approach and characterized the appearance and motion within these space time volumes by the computation of histograms of oriented gradient (HOG) and histograms of optical flow (HOF). Their results support the hypothesis of combining both feature types and similar suggestions have been made by other authors [28, 43, 47]. A recent evaluation of local spatio-temporal features for action recognition by Wang et al. [65] deduced that the combination of gradient-based and optical-flow-based descriptors is the best choice for such a task.

Hence, in our work we follow the idea of Mauthner et al. and use their method to represent the appearance and motion cues. We refer the reader to [43] for a more detailed description of their work. However, for completeness we review the relevant parts regarding feature extraction and emphasise towards our extensions afterwards.

HOG descriptors [13] have been proven to be effective in the context of person detection. In contrast to previous approaches [62], which utilized only appearance information, Mauthner et al. combined them with motion information derived from dense optical flow. Due to GPU-based flow estimation [75] both cues could be extracted in real time.

Appearance Features: For every frame $\mathbf{I}_t \in \mathbb{R}^{m \times n}$ at time t the gradient components $g_x(x, y)$ and $g_y(x, y)$ are computed for every pixel location (x, y) by simple filter operations. Both the magnitude $m(x, y)$ and the signed orientation $\Theta_S(x, y)$ are computed by

$$m(x, y) = \sqrt{g_x(x, y)^2 + g_y(x, y)^2} \tag{3.14}$$

$$\Theta_S(x, y) = \tan^{-1} \frac{g_y(x, y)}{g_x(x, y)}. \tag{3.15}$$

To ensure invariance according to the order of intensity changes, the sign of the orientation Θ_S is adapted as follows:

$$\Theta_U(x,y) = \begin{cases} \Theta_S(x,y) + \pi & \Theta_S(x,y) < 0 \\ \Theta_S(x,y) & \text{otherwise.} \end{cases} \tag{3.16}$$

To obtain the final HOG descriptor, each patch is structured into 10×10 non-overlapping cells. Each cell is represented by a histogram of 9 orientation bins and their counts are weighted by the magnitude at each pixel location. A grid of 2×2 cells is combined into overlapping blocks, normalized by using the L2-norm, and finally all blocks are concatenated into one vector.

Motion Features: The same principle is used to encode motion information, which provide an additional cue for the action recognition method. To estimate the dense flow for frame $\mathbf{I}_t$ two consecutive frames at time t and $t-1$ are required. Mauthner et al. used the public available GPU-based implementation [1] of the method proposed by [75]. The derived disparities $\mathbf{D}_t \in \mathbb{R}^{m \times n}$ in x and y direction are denoted as $d_x(x,y)$ and $d_y(x,y)$, respectively. They are used to compute the magnitude and orientation similar to the appearance features (see eq. (3.14) and eq. (3.15)). To be sensitive to the direction of the motion the signed orientation Θ_S is quantized into 8 bins. The same cell/block configuration as described above provides the final HOG descriptor for the motion cue.

For the rest of the work we used the terminology of HOG and HOF features to mention HOG representations of appearances and motion based on flow estimates, respectively.

Sparse Representation: Up to this point we applied the same procedure for feature extraction like Mauthner et al. [43]. Although they also used NMF to obtain a compact representation, we considered on a more sparse and part-based encoding of these features. This was achieved by using the NMF approach with sparseness constraints proposed by Hoyer [26] and a preceding structured initialization introduced by Boutsidis and Gallopoulos [9]. We refer the reader to section 2.2 for a formal description of these methods.

Briefly, during training all n-dimensional features vectors constituted an $n \times m$ matrix $\mathbf{V}$, where m denotes the number of features in the trainings set. The NMF algorithm with

[1] `http://gpu4vision.icg.tugraz.at`

sparseness constraints derived an approximate factorization of the form

$$\begin{aligned} \mathbf{V} &\approx \mathbf{WH} \\ \text{s.t.} \quad & 0 \leq \mathbf{W}, \mathbf{H} \\ & \text{sparseness}\,(\mathbf{w}_i) \quad = \quad S_w. \end{aligned}$$

Once such a factorization is found, we projected the test features using the multiplicative update rule eq. (2.2), while keeping the matrix of basis vectors $\mathbf{W}$ constant. This procedure was accomplished for HOG and HOF features individually. The resulting coefficients $\mathbf{h}_i^{HOG}$ and $\mathbf{h}_i^{HOF}$ for each image i are finally concatenated into one vector.

Hoyer's extension allows the user to specify the desired sparseness of $\mathbf{W}$ and $\mathbf{H}$. However, we limited only $\mathbf{W}$ to a level of $S_W = 0.7$. This choice was motivated by the findings in section 2.3. There we successfully demonstrated that such a representation is (a) more compact and (b) beneficial for the later classification stage, if the desired level exceeds the *natural* sparseness obtained by the NMF algorithm of Lee and Seung [35] (0.5 ± 0.03 and 0.5 ± 0.05 for HOG and HOF features, respectively). We limited the number of iterations to 150 for all experiments in this chapter, since we observed no significant improvement of the approximation error $\| \mathbf{V} - \mathbf{WH} \|^2$.

Note, for the rest of the work we used the term HOG/HOF features or short features as synonym for the representation obtained after the NMF projection step.

3.3.2 Datasets for Human Action Recognition

To achieve comparability among different algorithmic approaches it is crucial to have two issues in mind. First, using the same data is an essential prerequisite to compare the own performance with results published by other authors. The second important issue, sometimes neglected by authors, is the evaluation method. We focus our discussion on that in section 3.3.3. To fairly benchmark different methods one has to pay attention to use exactly the same input data and the same procedure for all of the algorithm under investigation.

Let us discuss the first issue – common data. In our work we used two standard datasets for human action recognition. Both of them are well known in this research field and many authors published their results on at least one of them. This helped us to directly rank different algorithms based on the measured performance and gave us the opportunity to compare a broader spectrum of approaches without reimplementing them.

Weizmann dataset: This standard benchmark for human action recognition, known as *Weizmann* human action dataset, was introduced by Gorelick et al. [22]. It contains 90 low resolution videos (180 × 144 pixel) with a rate of 50 deinterlaced frames per second. Ten different actions, namely running, walking, skip (jumping forward on on leg), jumping jack (denoted as jack), jump (jumping forward on two legs), pjump (jumping in place), gallop side ways (denoted as side), waving with one hand, waving with two hands, and bend, were performed by nine different subjects in front of a static background. Some of the actions are periodic like waving or jumping jack, while others are performed only once like e.g. walking, running or galloping side ways. Another property of the recorded data is that directed actions like walking etc. are not consistently performed from one side to the other, so several subjects walked from left to right while others walked into the opposite direction. Figure 3.5 shows illustrative examples for each of the actions performed by different subjects.

Figure 3.5: Illustrative examples from the Weizmann action dataset. Ten natural actions are performed by nine different subjects in front of a static background. Some videos contain multiple action cycles, while other sequences showing actions like 'run', 'side', or 'jump' from either left to right or in the opposite direction.

Unfortunately, there exist two 'versions' of this dataset, one containing ten actions as described above, and one subset that omits the action *skip*. In the literature both versions are referred as Weizmann dataset [38, 48], and as a consequence one has to pay attention when comparing reported performance results. For the entire dataset comprising ten different actions we refer to the work by Gorelick et al. [22] and for the subset to Blank et al. [8] as stated on the authors page[2].

[2]see `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.htm`

KTH action dataset The second large collection of public available video sequences containing human actions is the KTH dataset and was introduced by Schuldt et al. [59]. The entire dataset is divided into four different recording settings ($s1 \ldots s4$). The first three settings are recorded outdoor with the following variations, constant size of the human during action performance ($s1$), scale changes via camera zoom or subjects movement ($s2$), and appearance variations caused by different clothes ($s3$). The last subset ($s4$) was recorded indoor in front of a homogeneous background with lighting variations. Each of the subsets consist of six types of human actions, namely boxing, hand clapping, hand waving, jogging, running, and walking performed by 25 individual subjects. The videos are down-sampled to a spatial resolution of 160×120 pixels and are recorded with an almost static camera with 25 fps. In contrast to the Weizmann dataset, where actions like walking and running are performed only once in every video, here all actions are performed multiple times in one of the 600 videos. Schuldt et al. provided additional information to segment those videos into 2391 sequences containing only one action cycle. While some authors [2, 33, 58] utilize these segmentation during training and testing, others do not rely on such an information [48, 55]. The KTH dataset as well as the additional segmentation information can be downloaded from the authors web page[3].

The wide acceptance of these two datasets in the research community as well as the public accessibility led us to the decision to use them in our study. Both datasets have their challenges, a larger number of different actions in the case of the Weizmann videos, as well as a large intra-class variations through different scenarios and many subjects for the KTH data.

The authors of this work are aware of other public available benchmark videos in the context of human action recognition, namely the SDHA high-level human interaction challenge [56], the INRIA XMAS dataset [70], the UCF sports actions [53], or the Hollywood human action dataset [33] to mention the most prominent. Other related communities, e.g., in the context of semantic analysis or event detection like TREC Video Retrieval Evaluation (TRECVID) [61] provide annotated video data as well. However, the large number of already published results influenced our decision towards Weizmann and KTH. Both of them allowed us to directly compare previous approaches with those presented in this work and results on that issue are discussed in section 3.4.3 and section 3.4.6 respectively.

Common data is an important issue to rank different approaches against each other.

[3] see `http://www.nada.kth.se/cvap/actions`

Figure 3.6: Example images from sequences in the KTH dataset. Each of the four rows ($s1 \ldots s4$) shows one of the six different actions performed in different scenarios. The first three rows are outdoor settings with varying difficulties, fixed scale (s1), with scale changes (s2), subjects wearing different clothes (s3), and the fourth row illustrates an indoor scenario (s4) with homogeneous background.

Another critical topic is a consistent and transparent evaluation procedure. We continue our discussion on this topic in the following section.

3.3.3 Evaluation Method

This section gives an overview of the evaluation schema and the parameters we used in the following experiments.

Every evaluation in this section was performed using a *leave-one-out-actor* setting. For each action one classifier was learned using all videos except those corresponding to the actor left out for testing. In a MIL learning schema samples are organized in bags, i.e., a unordered set of samples. In the context of HAR, such a bag comprises image regions (encoded using HOG and HOF features) from all frames of an specific action sequence. The label of the bag is derived from the action performed in that video. We used multiple detection windows around a perfect annotated sample for each frame, to generate negative instances [4]. We describe the training and test procedure in the following.

[4]The process of feature extraction was implemented by our colleague Thomas Mauthner, who provided the raw HOF and HOG descriptors.

Training stage: All training instances regardless of the bag membership were used to derive a structured initialization for the subsequent matrix factorization. This was achieved by applying the NNDSVD algorithm. The obtained matrices **W** and **H** as well as all trainings data were fed into the subsequent sparse NMF approach, and the derived feature representations constitute the input of each learning algorithm. Following the leave-one-out-actor cross validation, all instances are organized in positive or negative labelled bags. Since, all MIL approaches used in our study assume a binary classification problem, we trained one classifier per action. The derived classifier is then tested on unseen action sequences.

Test stage: First, we projected the raw HOG and HOF descriptors of all test videos into the previously obtained feature space. This was achieved using the multiplicative update rule eq. (2.2), while keeping the matrix of basis vectors **W** (derived during training) constant. Second, these features were again organized in bags, and these bags were used to test each individual classifier. Finally, the results of all action classifiers where combined using a *winner take all* strategy to obtain the final decision regarding the action label.

In comparison to other authors [2, 27, 62] it is important to note that we did not perform any kind of preprocessing like background subtraction, down-sampling or manual selection of action cycles.

Decisions concerning the dimensionality of the feature space are crucial. As already mentioned in section 3.3.1, we used the same feature extraction methods as Mauthner et al. [43]. They suggested 80-100 basis vectors for the task of action recognition on the Weizmann dataset and reported no significant improvement regarding classification performance using a higher dimensional feature representation. Since, our feature encoding only differ in the method of matrix factorization (unconstrained versus sparse NMF), we generalized their findings to our recognition task. However, to roughly examine the influence of the feature dimension, we evaluated a few experiment on that issue. In order to limit the computational effort, we compared the classification accuracy of a single MIL approach (MILES) on the Weizmann dataset. Using 80, 160 and 250 modes for HOG and HOF representations, we achieved 98,8%, 97,78% and 97,22% repectively. Since, our objective concernes a comparative study of different MIL classifiers, it is important that all algorithms share exactly the same input data. Therefore, we fixed the number of basis vectors $r = 80$ for both, HOG and HOF features, in all subsequent experiments.

We complete this section with a summary of the parameter settings used for each MIL algorithm in our comparative study.

- For the method denoted as *Kinematic-MIL*, we used four principal dimensions for HOG and HOF features individually. This was suggested by Ali and Shah [2] and the resulting dimension of $\mathbb{F}_C$ is then '$4 \times 2 \times$ number of bags', since we treated appearance and motion descriptors as different kind of features. During kernel embedding (see eq. (3.6)) we fixed σ to 0.7, similar to Ali and Shah.

- We used the following parameters suggested by Zhou et al. [77] for the *miGraph* algorithm. Within eq. (3.8) the threshold δ_i was set to the average distance between instances in bag $\mathbf{B}_i$ and the value of σ was set to 1.0. The SVMs used Gaussian RBF kernels with $C = 100$ and $\gamma = 5$.

- Three parameters need to be specified for *MILES*. The value of μ was set to $^1/_8$ for the Weizmann and to $^1/_{24}$ for KTH benchmark, to equally penalize errors in the binary classification problem. The parameters λ and σ were determined in a cross-validation step on the training set. We found that $\lambda = 0.5$ and $\sigma = 2.5$ performed best for the Weizmann database. Since we used the same kind of features for both benchmark datasets, we kept these parameters for all subsequent experiments.

- The results for *MI-SVM* are based on SVMs with linear kernels. The choice was motivated by experimental results on image data in the original work of Andrews et al. [3].

- The last MIL approach (*Citation-NN*) in our study needs two parameters to be specified. We followed the standard parameters suggested by Wang and Zucker [66] and fixed the number of citations to $C = R + 2$, with $R = 2$ (number of references).

3.4 Experimental Results

In this section we present results from our experiments, explain the performance evaluation and discuss about the consequences regarding potential applications. Since we stress the terminology of HOG and HOF, we want to remind the reader that HOG and HOF features denotes a histogram of X representation, with gradients to encode appearances in the case of HOG and motion features based on flow estimates for HOF, respectively. Moreover, we used the term HOG/HOF features or short features as synonym for the representation obtained after the NMF projection step.

3.4.1 Static and dynamic features for Action Recognition

The first experiment was designed to investigate the influence of static and dynamic features with respect to the accuracy of the classifier. Wang et al. [65] claimed that a combination of gradient and optical flow based descriptors work best in the context of HAR. Similar results have been suggested by other authors [27, 33, 58]. Mauthner et al. [43] already showed that their proposed combination of HOG and HOF features improves the overall recognition accuracy. Since we used their method for extracting features in the first stage, but in contrast investigated towards a sparse representations in the later stage. This evaluation was intended to demonstrate whether the claimed conclusions hold for our specific setting.

The Weizmann dataset [22] was selected to demonstrate the effect of pre-processing the data. As described in section 3.3.2, the video sequences for dynamic actions (e.g., walking, running, jumping) are not consistently performed in both directions. We conjecture that this specific property leads to additional difficulties for the final classifier. In order to compensate for that, all videos are mirrored across the vertical axis, to obtain both versions of the action sequence.

Table 3.2 depicts the performance results using the Weizmann dataset for three different representation, static only (HOG), dynamic only (HOF) and a combination of both. MILES was used for MIL and the accuracy was evaluated in a leave-one-out-actor setting. Other classifiers could have been used to demonstrate the effect, however we used one of the methods introduced in section 3.2, since we compared their recognition rate in section 3.4.2. The second and third column of table 3.2 denotes the feature type, while the fourth column specifies whether the mirrored versions of the action sequences have been used in addition to the original data. The fifth column indicates the number of false classifications and in brackets the total number of videos (denoted as bags in the context of MIL). Finally, the average accuracy is given in the rightmost column.

The rows a – c of table 3.2 show the performance when the pre-processing described above was enabled, while rows d – f show results for the dataset without any modification of the original data. In both groups the combination of static and dynamic features works best and we achieved an accuracy of 98,8% and 95,5%, respectively. This supports the findings of other authors [33, 58, 65] and demonstrates that a sparse representation in combination with MIL is able to achieve highly competitive recognition results.

When comparing equal feature combinations from both groups it can be shown that the pre-processing step beneficially affects the accuracy of the classifier by $\sim$ 3%. In-

	HOG	HOF	flipped	wrong (total) bags	accuracy
a	X	X	X	2 (180)	**98,8%**
b	X		X	14 (180)	92,2%
c		X	X	10 (180)	94,4%
d	X	X		4 (90)	**95,5%**
e	X			10 (90)	88,8%
f		X		18 (90)	80,0%

Table 3.2: Combining static (HOG) and dynamic (HOF) features improves the overall performance. Here MILES was used as an MIL approach on the Weizmann dataset. An extra performance gain of approx. 3% can be achieved by additionally using the mirrored versions of the action sequences.

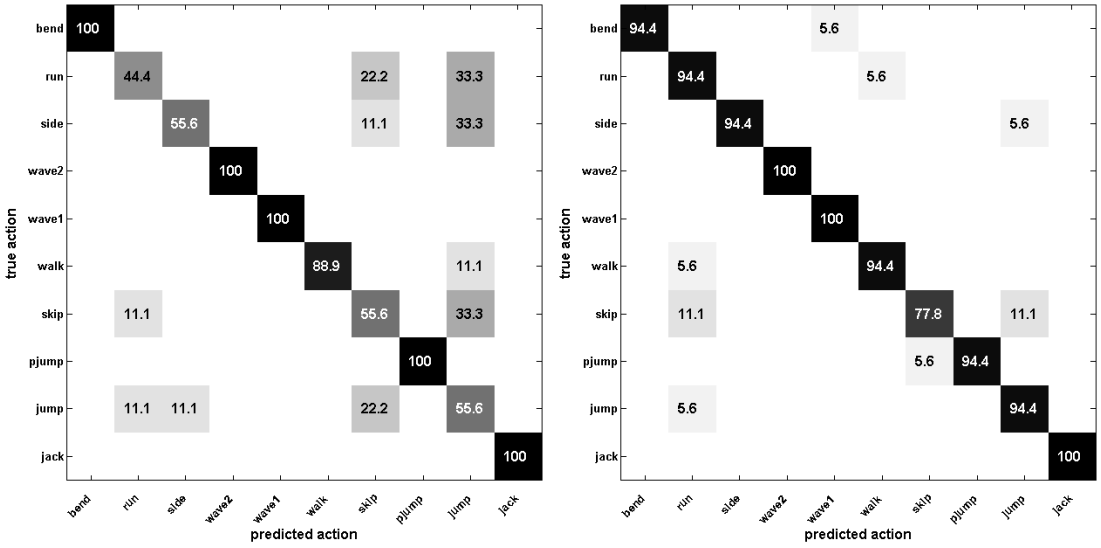
terestingly, no general statement can be made in favour of one particular feature type. While rows e and f depict that HOGs perform better compared to HOFs, reciprocal conclusions have to be made when comparing rows b and c in table 3.2. Furthermore, when comparing the experiments where motion features alone represent the action sequences (rows c and f), a performance gain of $\sim 14\%$ can be achieved when the mirrored sequences complement the original data. This finding is not surprising, since dynamic features like HOF have to be sensitive to the direction of the motion. Therefore, an action like walking from left to right is differently represented compared to walking in the opposite direction. The resulting HOFs are varying and the classifier is not able to generalize accross these differences. Figure 3.7 illustrates this difficulty by comparing the confusion matrices. In both experiments we only used motion features (HOF) and the MILES approach as a MIL classifier.

In the next section we compare the performance results of each individual MIL approach introduced in section 3.2.

3.4.2 Performance on the Weizmann Dataset

Within this section we present the result of our comparative study and address the following question: Is MIL an appropriate learning schema for an action recognition task? Therefore we evaluated five different frameworks for MIL and compared the results against a linear SVM classifier.

Table 3.3(a) summarizes the classification performance from all MIL approaches used in our study. We investigated each individual method using a leave-one-subject-out evaluation schema. From the previous experiment we conclude that the combination of appear-



(a) When using the original action sequences we get a recognition rate of 80%. The major confusion occurs between directed actions.

(b) An increased accuracy of $\sim$ 94% can be achieved when the mirrored versions of each action sequence are used in addition to the original data.

Figure 3.7: Confusion matrix of the MILES classifier expressed as a percentage. For both experiments only motion features (HOF) have been used (see table 3.2 row c and f). Wrong classifications occur more likely for directed actions e.g. 'skip', 'run', or 'jump' due to the sensitivity of the HOF features. This effect could be minimized, when the mirrored versions of the sequences complement the original dataset.

ance and motion features work best for the Weizmann dataset. Therefore we calculated both HOG and HOF representations and derived the sparse encoding of these by applying the NMF algorithm described in chapter 2. Details for each individual MIL method can be found in section 3.2. Furthermore, the proposed standard parameter set was used for each algorithm.

As a reference method a linear one-versus-all SVM was used as baseline classifier. This algorithm is denoted by the acronym *SVM baseline* in table 3.3(a) as well as in the following experiments. We used the public available *LIBSVM* [5] implementation for this experiment. The same algorithm was used by Mauthner et al. [43] but in contrast to them, we do not focus on instance action recognition and therefore report results obtained by evaluating the entire action sequence. Thus, the majority votes decide on the predicted action label for every test video.

[5] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

(a) Recognition performance.

algorithm	accuracy
Citation-NN	93,3%
MI-SVM	86,6%
MILES	**98,8%**
miGraph	97,8%
Kinematic-MIL	97,8%
SVM baseline	96,7%

(b) Timing comparison.

algorithm	training	test	total
Citation-kNN			664,2 sec.
MI-SVM			451,8 sec.
MILES	24,1 sec.	**35,1 msec.**	
miGraph	27,8 sec.	3,4 sec.	
Kinematic-MIL	7,9 sec.	1,1 sec.	
SVM baseline	53,3 sec.	1.3 sec.	

Table 3.3: Experimental results for the Weizmann action dataset [22].

From table 3.3(a) we can see that three out of five methods achieved a better classification result compared to the baseline supervised learning approach. With the exception of *MI-SVM*, we were able to report state of the art recognition accuracy for this dataset. See section 3.4.3 for a more detailed comparison with other approaches. Little confusion occurred between running and the action denoted as 'skip' (jumping forward on one leg), independent from the used classifier. In the case of *MILES*, two video sequences showing the 'skip' action are classified as running action. These two misclassified videos led to a final accuracy of 98.8%. The observation corresponds to results reported by other authors [2, 48, 55, 62].

However, the outcome of this experiment is significant because it demonstrates an important feature that motivated our work. Thus, we conclude that MIL provides an alternative learning schema in the context of human action recognition.

Some applications require methods that respond within a specified period of time. Table 3.3(b) summarizes the duration of the training and testing stage. We report the average over all nine runs of the cross validation process. Within each single run ten different classifiers were learned on 160 trainings sequences and evaluated on 20 test videos in a one-vs-all configuration. All experiments were performed on a 3 GHz Intel© Core™ 2 Duo processor with 3 GB RAM. Although all of them have been implemented in a Matlab© framework, the results are not purely quantitative. Some methods use optimized code while others might suffer from an inefficient implementation. Note that the library used for *Citation-NN* and *MI-SVM* did not support a separate logging of trainings and testing. Due to this difficulty, we reported the total time spent in both stages.

The results in table 3.3(b) clearly show that *MILES* provides the most efficient method for classifying new action sequences. This finding confirms one of the major advantage of

this approach (see section 3.2.3). The combination of 98.8% classification accuracy and the fast response time ($\sim 35ms$) suggests MILES as a promising approach for practical applications.

Weizmann - nine action subset: As already mentioned in section 3.3.2, some authors [8, 38, 58] reported evaluation results on an earlier version of this benchmark. The subset omits the jumping-forward-on-one-leg action denoted as 'skip', and comprises therefore nine different actions from nine individual subjects. We evaluated all five MIL approaches as well as the baseline SVM classifier using this dataset and table 3.4 summarizes the results. Without the 'skip' action, no further confusion occurred between directed actions and some algorithms were able to recognize all sequences without failure.

	Citation-NN	MI-SVM	MILES	miGraph	Kinematic-MIL	SVM baseline
accuracy	96,3%	98,1%	100%	100%	98,8%	100%

Table 3.4: Classification performance on an earlier version of the Weizmann action dataset [8]. This subset omits the action 'skip' and typically better recognition rates are achieved.

3.4.3 Comparison to previous papers

We summarize various performance results from other authors within this section to complete the discussion on the Weizmann dataset. The ranking in table 3.5 reviews the more prominent approaches since 2005, without intending to be exhaustive. As mentioned in section 3.3.2 some authors did not explicitly mention which version of the dataset they used for their evaluation. Moreover, there are a few comparisons that mixed up results from both datasets. Table 3.5 shows the accuracies achieved for the ten action benchmark [22] in column four. Results for an earlier version [8] containing nine actions are given in column five. Both results are given where available. The approaches are sorted according to their date of publication and the first row depicts the best performing MIL approach out of our comparative study.

As can be seen most of the authors evaluated their algorithms using a leave-one-subject-out cross validation schema, except Jhuang et al. [27] and Junejo [28]. The latter used a different version of the nine action dataset, where they omitted the action 'wave2' (waving with two hands) instead of the 'skip' action.

The conclusion from table 3.5 is that the recognition performance on this benchmark is already saturated and perfect accuracy has been reported by [22, 36, 42]. The sparse

author		year	evaluation	accuracy	
				10 action [22]	9 action [8]
MILES (MIL)	[12]	2011	loo-cv	98,8%	100 %
Mauthner et al.	[42]	2010	loo-cv	**100 %**	100 %
Ali & Shah (MIL)	[2]	2010	loo-cv	95,8%	–
Yao et al.	[72]	2010	loo-cv	95,6%	–
Bregonzio et al.	[10]	2009	loo-cv	96,7%	–
Mauthner et al.	[43]	2009	loo-cv	88,9%	–
Yeffet & Wolf	[74]	2009	loo-cv	–	100 %
Lin et al.	[36]	2009	loo-cv	**100 %**	–
Roth et al.	[55]	2009	loo-cv	94,2%	97,0%
Thurau & Hlaváč	[62]	2008	loo-cv	94,4%	–
Liu et al.	[38]	2008	loo-cv	–	90,4%
Schindler & van Gool	[58]	2008	loo-cv	–	99,6%
Junejo et al.	[28]	2008	n-fold cv	–	95,3%
Niebles et al.	[48]	2009	loo-cv	90,0%	–
Niebles & Fei Fei	[47]	2007	loo-cv	–	72,8%
Jhuang et al.	[27]	2007	split	–	98,8
Gorelick et al.	[22]	2007	loo-cv	**100 %**	–
Blank et al.	[8]	2005	loo-cv	90,0%	–

Table 3.5: Action recognition precision on the Weizmann benchmark. The results of other authors are copied from their paper (see references). The approaches are sorted by their date of publication and the best MIL method (MILES) is shown on top. We differentiate between two versions of the dataset, the nine action subset and the ten action benchmark. Performances for both variants are given in the fourth or/and fifth column. The evaluation method is shown in column three.

feature representation proposed in chapter 2 in combination with a MIL approach achieved competitive recognition precision for three out of five MIL schemas. We demonstrated that these methods denoted as *MILES*, *miGraph* and *Kinematic-MIL* achieved $98,8\%$, $97,8\%$ and $97,8\%$ on the ten action benchmark. The comparison with other results in table 3.5 indicate MIL as a promising alternative in the context of human action recognition.

3.4.4 Performance on the KTH Dataset

In section 3.4.2 we demonstrated that MIL approaches are able to achieve competitive performance results on the Weizmann benchmark. In this section we report results for the KTH dataset described in section 3.3.2 and again compare five MIL methods against a linear SVM classifier as a baseline recognition framework. We evaluated each subset separately using a leave-one-subject-out strategy as described in section 3.3.3 and calculated

the overall precision by averaging across all subsets.

Table 3.6 shows the results for each KTH subset (s1,...,s4) individually and summarizes the average accuracy in column six. The algorithms in this table are sorted according to their overall performance, with the best one on top. The bottom line comprises the data for a linear SVM classifier and is denoted as *SVM baseline*. For this algorithm every frame has to be classified and the predicted action label for one video is computed by a majority voting. All other methods handle each video as bag of instances and inherently derive the recognized action label.

As can be seen in table 3.6 the MIL approach denoted as *miGraph* achieved $93,3\%$ accuracy and performed slightly better compared to the baseline. The method proposed by Chen et al. [12] (MILES) is the second best MIL classifier but lost $\sim 1\%$ compared to the baseline SVM. This result is contrary to the experiments on the Weizmann benchmark, where MILES outperformed all other methods in the evaluation.

When comparing both datasets it seems obvious that the KTH is the more challenging one, with each individual subset (s1,...,s4) having its own difficulties (see section 3.3.2 for a detailed description). From our results in table 3.6 and those reported by other authors [27, 37, 58] we conclude as follows.

algorithm	accuracy on subset				avg. accuracy
	s1	s2	s3	s4	
miGraph	96,3%	87,3%	92,3%	97,3%	93,3%
MILES	94,0%	85,3%	93,3%	94,7%	91,8%
Kinematic-MIL	92,3%	84,7%	86,3%	87,0%	87,6%
MI-SVM	89,7%	79,7%	84,7%	86,0%	85,0%
Citation-NN	84,3%	69,3%	83,3%	87,3%	81,2%
SVM baseline	97,0%	86,0%	93,0%	94,6%	92,7%

Table 3.6: Performance comparison on the KTH benchmark. Each subset was evaluated individually using a leave-one-subject-out schema and the overall accuracy was calculated by averaging across all portions. The MIL algorithms are sorted according to the average recognition rate and the bottom line comprises results for a linear SVM classifier.

We derive a ranking of the individual recoding settings (s1,...,s4) based on the results in table 3.6. The outdoor subset with constant human scale (s1) as well as the indoor setting with varying illumination conditions (s4) share a similar level of complexity. Subset s3 (outdoor scenario with different clothes) is more challenging due to appearance changes and the most difficult subset is s2 (outdoor with scale changes). While miGraph and Citation-NN achieved their best result for subset s4, Kinematic-MIL, MI-SVM and

the baseline SVM approach performed best for subset s1. Since all algorithm share the features representation, this variation originates from the different classifiers. However, it demonstrates that our representation is able to handle illumination changes very well. A fact we expected from the appearance features (HOGs).

The common observation across all classifiers within our study is that the used representation is not able to model changes in the scale of the human body. This was obviously predictable, since the simple but efficient principle of histogram based features was never designed to handle such variations. Note that the performance differences between the best subset and s2 is larger than 10%.

Figure 3.8 depicts the confusion matrix for the best MIL approach on the KTH benchmark. The algorithm proposed by Zhou et al. [77] outperformed the other MIL methods for all subsets, except for s3 where MILES was best. The confusion matrix illustrates the main difficulties for this dataset. The vast majority of errors occur within two groups of actions, hand-based (boxing, clapping and waving) and leg-based (jogging, running and walking). One can imagine that these actions share common poses and motions with differences in order and speed. Consequently, the resulting features are similar and the classifiers are not able to differentiate between them. A detailed view on fig. 3.8 shows that most mistakes happen between clapping and waving according to hand-based actions, while jogging and running are more likely to be misclassified inside the second group of actions. These results are evident with those reported by other authors like [21, 36, 74].

Our analysis of different MIL approaches on the KTH dataset and the comparison with a standard linear SVM classifier raises the following question: Are the investigated MIL methods robust against noisy features and therefore applicable in a typical human action recognition scenario? We consider this issue in the following section.

3.4.5 Robustness to Noisy Patches

Investigating the robustness of methods is essential for practical applications. However, it is crucial to define what kind of corruption the experiment investigates. Therefore, it is important to emphasize that we do not focus on images that are degraded by, e.g., compression algorithms, quantization errors or pixel noise caused during capturing or transmission of the video stream. Instead, we assume that the feature extraction process is capable to deal with this sort of degradations.

Contradictory, we address another, from our point of view, more ordinary source of noise: an imperfect person detector. To simulate the localization errors of the person

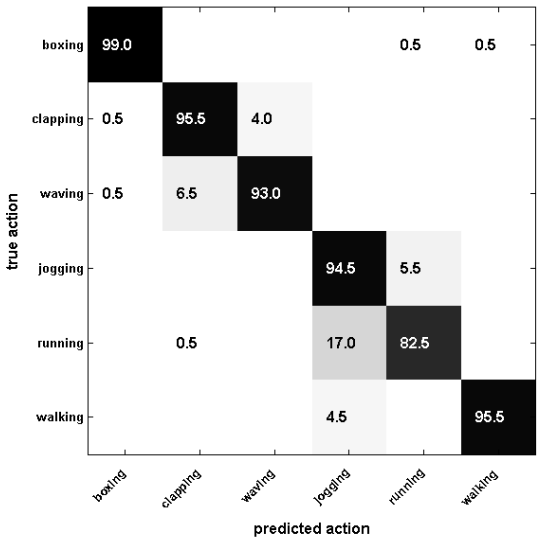


Figure 3.8: Overall confusion matrix for the miGraph [77] approach on the KTH dataset. The majority of errors occur between jogging and running as well as hand clapping and waving, while most other actions are easily distinguished.

detector, we replaced $\sim 22\%$ of the perfect aligned instances with their noisy counterparts. This means for each individual recording setting (s1,...,s4) a subset of $\sim 20\%$ randomly choosen frames was selected and processed as follows. The perfect aligned bounding box (BB) of the person detector was shifted away from its initial position and the feature extraction process was repeated on the resulting region of interest. The generated noisy feature substitutes the original one in the corresponding action sequence.

Figure 3.9 illustrates three different levels of noisy feature generation. The localization error was simulated by shifting the original BB, shown on the left side and denoted as *no noise* in the figure as well as in table 3.7. Column two depicts the first noise level (*quarter noise*), where the BB is 25% away from its original position. *Half noise* denotes 50% overlap with the original BB and the right side of fig. 3.9 shows example background positions to generate an *off noise* feature.

The resulting datasets were used during training and testing, while following the leave-one-out-actor evaluation schema described in section 3.3.3. Table 3.7 summarizes the overall accuracies for every MIL algorithm by averaging across all subsets (s1,...,s4). The bottom line shows the performance of the baseline SVM approach. When comparing the best MIL approach (first row) with the linear SVM baseline (bottom row) it can be seen that both methods perform almost equally with a tolerance of $\sim 1\%$ for the first two noise

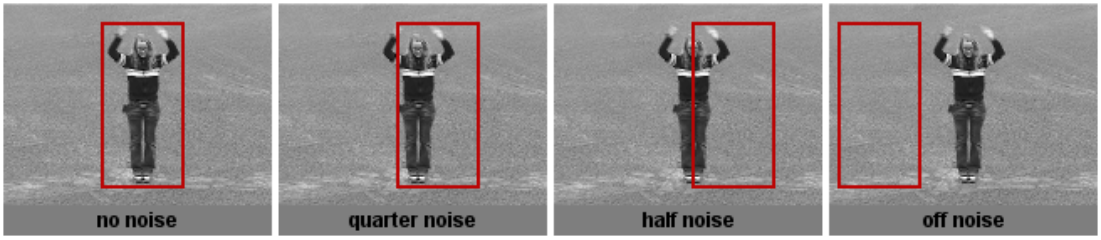


Figure 3.9: Schematic illustration of simulated detector errors on the KTH dataset. From left to right: original bounding box (BB) superimposed on an example frame, the BB shifted 25% away from its position, 'half noise' denotes a 50% shift, background patches are used to generate 'off noise' features.

levels and within $\sim 2\%$ for the background noise level shown in the rightmost column of table 3.7.

One conclusion that might be drawn from this is, that there is no clear indication that MIL is superior to a supervised learning method. This experiment regarding robustness and the results from the previous section have shown that only one of five MIL methods achieve equal or better accuracies compared the the baseline SVM approach.

The best performing algorithm proposed by Zhou et al. [77] and denoted as *miGraph* utilizes the relations between images of one action sequence, assuming that the structure in subsequent frames are not independently and identical distributed. This method demonstrates the significance of the inherent structure within human actions and has shown its advantages on more complex datasets.

The second best MIL approach called MILES [12] is based on a diverse density framework [14]. The method proposed by Chen et al. is very fast in the test phase since it uses only very few features to embed every video into a single instance space. However, this

algorithm	avg. accuracy			
	no noise	quarter noise	half noise	off noise
miGraph	93,3%	92,3%	92,1%	80,6%
MILES	91,8%	89,9%	88,4%	81,6%
Kinematic-MIL	87,6%	84,5%	82,4%	76,3%
MI-SVM	85,0%	79,3%	76,7%	75,0%
Citation-NN	81,2%	78,4%	73,1%	74,3%
SVM baseline	92,7%	90,9%	91,2%	82,4%

Table 3.7: Performance comparison with different levels of detector noise for the KTH action dataset. On average 22% of all frames are replaced by their degraded counterparts. The average accuracy across all four subsets is reported for each individual classifier.

might be one reason why this algorithm is not able to outperform the baseline SVM.

Across all experiments Kinematic-MIL, a method proposed by Ali and Shah [2], is the third top multiple instance learning schema. We want to emphasize that our framework with sparse representations of both HOG and HOF features is superior to the original work from Ali and Shah. Contrary to their kinematic features, we achieved an increased performance of +2% for the Weizmann dataset and attained competitive recognition results for the KTH benchmark. Note, contrary to their framework, we do not rely on any preprocessing, scaling or additional segmentation information regarding single action cycles.

As already mentioned in section 3.4.4 the final decision in the case of the linear SVM classifier is derived by a majority voting across all images in a test sequence. Thus, all frames have to be evaluated individually. We suspect that the final voting schema is beneficial since a single-frame based evaluation leads to a accuracy of $\sim 83\%$ for the KTH dataset without noise.

We showed a detailed analysis of five MIL approaches on the KTH dataset and investigated their robustness against noise. The comparison against a linear SVM classifier suggests that the best performing MIL approaches achieve at least equal recognition results. Moreover, the empirical evaluation in table 3.7 depicts that the average accuracy of all methods (except for Citation-NN) decreased by $\geq 10\%$ when we simulate $\approx 22\%$ false detections.

3.4.6 Comparison to previous papers

Without claiming to be exhaustive, we summarized the most prominent performance results on the KTH dataset in table 3.8. The first row in this table presents the accuracy achieved by the best performing MIL algorithm out of our comparison experiment. The results from other authors are sorted according to their date of publication.

In general, one has to be careful when comparing performances between different approaches. Especially for this dataset we noticed diverse variations in the evaluation method. Schindler & van Gool [58] for example used a 5-fold cross validation and reported the average accuracy over 5 runs, while Yeffet & Wolf [74] and Kovashka & Graumann [31] used the protocol of Schüldt [59], who made this dataset public available. Column three in table 3.8 addresses this issue and one can see that the evaluation method is evenly distributed between leave-one-subject-out ('loo-cv') and n-fold cross validation (denoted as 'split').

Another difficulty concerning performance comparison arises from the following fact. The KTH has been treated either as four distinct subsets (s1,...,s4) as we did, or as one large dataset with strong intra-subject variations. Column five summarizes the first case, where each subset has been evaluated separately and the average precision was reported and the rightmost column shows results from authors, who handled the KTH as one large dataset. We emphasised the best approaches in both columns.

In can be seen in table 3.8 that the best performing methods are either based on a prototypical representation or they used a hierarchy of feature combinations. Mauthner et al. [42] for example proposed an efficient prototype-based encoding using four distinct feature cues together with a temporal weighting between these channels. Lin et al. [36] in contrast established a frame-to-prototype correspondence of the action and used dynamic time warping (DTW) to automatically identify optimal matching segments between the

author		year	evaluation	precision	
				average	all in one
miGraph (MIL)	[77]	2011	loo-cv	93,3%	–
Yao et al.	[72]	2010	loo-cv	–	92,0%
Kovashka & Grauman	[31]	2010	split	–	**94,5%**
Ali & Shah (MIL)	[2]	2010	split	–	87,7%
Mauthner et al.	[42]	2010	loo-cv	**97,4%**	–
Yao & Zhu	[73]	2009	split	88,0%	87,8%
Lin et al.	[36]	2009	loo-cv	95,8%	93,4%
Yeffet & Wolf	[74]	2009	split	–	90,1%
Bregonzio et al.	[10]	2009	loo-cv	–	93,2%
Gilbert et al.	[21]	2009	loo-cv	96,7%	–
Gilbert et al.	[21]	2009	split	94,5%	–
Schindler & van Gool	[58]	2008	split	90.7%	92,7%
Niebles et al.	[48]	2008	loo-cv	–	83,3%
Mikolajczyk & Uemura	[45]	2008	loo-cv	93,2%	–
Jhuang et al.	[27]	2007	split	91,7%	–
Nowozin et al.	[49]	2007	split	–	87,0%
Dollár et al.	[15]	2005	loo-cv	–	81,2%
Schüldt et al.	[59]	2004	split	–	71,7%

Table 3.8: Recognition performance on the KTH benchmark. The results of other authors are copied from their paper (see references). The best MIL method from our comparative study is listed first, while all other authors are sorted according to their publication date. Column three indicates the evaluation method, column four depicts the average accuracies over all four subsets of the KTH and the rightmost column summarizes the accuracy for an all-in-one evaluation.

training and a test sequence. Gilbert et al. [21] as well as Kovashka & Grauman [31] learned a hierarchy of spatio-temporal feature configurations. The reoccurring pattern in [21] are distinct compared to other actions and support the later classification stage. In [31] local neighbourhood combination of space-time interest points and their most discriminative descriptors w.r.t. an action category are combined to a vocabular hierarchy.

Given a choice between two theories, take the one which is funnier.

Blore's Razor

Chapter 4

Conclusion

4.1 Summary

We investigated a sparse, part-based representation obtained by a constrained NMF approach. The choice of NMF is based on the assumption, that complex actions could be composed of multiple limb movements. We encoded appearance and motion cues from the entire human body to derive a suitable representation of actions. The additive nature of NMF relaxes the holistic characteristic of the feature extraction and we demonstrated that such an representation beneficially supports the consecutive classification process.

In chapter 2 we analysed the constrained NMF approach proposed by Hoyer [26] and demonstrated the importance of a structured initialization for this factorization. The correlation of sparseness and feature dimensionality is discussed and we evaluated the obtained representation in the context of face recognition. The results of our experiments support our basic assumption, that a sparse, part-based representation is able to increase the performance of the subsequent classifier.

We used the ORL database of faces to evaluate two different classifiers. In addition to a simple NN classifier and we applied the LMNN approach proposed by Weinberger et al. [69]. In both cases, we achieved an increased accuracy of $\sim 1.25\%$, if the sparseness constraint was choosen appropriately. This means that the representation has to be suffiently rich to be discriminative. First, the dimensionality of the feature space has to be large enough and second, the sparsity of the encoding must exceed the *natural* sparseness obtained by an unconstrained NMF. In our experiment we empirically demonstrated this, using e.g. 50 modes and a level sparseness level of $S_W = 0.8$ and $S_W = 0.7$ for NN classifier and the LMNN approach, respectively.

The main topic of our work is a comparative study of different MIL approaches in the context of HAR. We investigated whether the MIL concept is suitable for an action recognition task and performed a thoroughly and consistent study using two state-of-the-art datasets. We emphasized the importance of clear structured experimental setup in order to achieve comparable results and evaluated five different MIL approaches.

Our representation of human actions is based on the appearance and motion features proposed by Mauthner et al. [43]. We extended their encoding by introducing sparseness constraints and evaluated the combinations of both cues. Moreover, we investigated the effect of video pre-processing on the Weizmann dataset. The video sequences for directed actions (e.g., walking, running or jumping) are not consistently performed in both directions. To compensate for that, all videos are mirrored across the vertical axis. When comparing the recognition accuracy obtained using motion features only, we observed an additional performance gain of $\sim 14\%$, using the described pre-processing step.

We systematically evaluated five MIL approaches on the Weizmann dataset. Three out of five methods, MILES $(98,8\%)$, miGraph $(97,8\%)$, and Kinematic-MIL $(97,8\%)$ achieved competitive or better results compared the our baseline – a linear SVM classifier $(96,7\%)$.

Additionally, we performed the same analysis on four different subsets of the KTH dataset. We evaluated each subset separately and reported the overall precision. In this case, the MIL method denoted as miGraph $(93,3\%)$ performed slightly better compared the baseline SVM classifier $(92,7\%)$. Since robust methods are essential for practical applications, we simulated an imperfect person detector, to generate misaligned samples. Up to a moderate noise level, the best performing MIL approach (miGraph) achieved equal or better performance results compared to the baseline (see table 3.7).

We summarized the accuracies of the most prominent approaches by other authors for both benchmark datasets and emphasized the differences in the evaluation process. It turned out that the best performing methods utilized either a prototypical representation or a hierarchy of feature combinations. However, when comparing the complexity of the superior approaches, it surprised us, that the proposed sparse part-based feature representation in combination with the MIL principle achieved competitive accuracies in both cases.

4.2 Summary of our Contribution

The main contributions of the thesis are:

Sparse Features We investigated a sparse NMF approach and illustrated the importance of a structured initialization towards a part-based representation. We demonstrated that such a representation is able to improve the consecutive classification process.

Comparative Study We considered the concept of MIL in the context of action recognition and performed a thoroughly analysis of five different MIL approaches using two state-of-the-art action datasets.

Bibliography

[1] Agarwal, S., Awan, A., and Roth, D. (2004). Learning to Detect Objects in Images via a Sparse, Part-Based Representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 26(11):1475–1490.

[2] Ali, S. and Shah, M. (2010). Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32:288–303.

[3] Andrews, S., Tsochantaridis, I., and Hofmann, T. (2003). Support Vector Machines for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*, pages 561–568.

[4] Babenko, B. (2008). Multiple Instance Learning: Algorithms and Applications. Technical report, UCSD Computer Science and Engineering Department.

[5] Babenko, B., Dollár, P., Tu, Z., and Belongie, S. (2008). Simultaneous Learning and Alignment: Multi-Instance and Multi-Pose Learning. In *Proc. of ECCV Workshop on Faces in Real-Life Images*.

[6] Babenko, B., Yang, M.-H., and Belongie, S. (2009). Visual Tracking with Online Multiple Instance Learning. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 983–990.

[7] Biederman, I. (1987). Recognition-by-Components: A Theory of Human Image Understanding. *Psychological Review*, 94(2):115–147.

[8] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as Space-Time Shapes. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 1395–1402.

[9] Boutsidisa, C. and Gallopoulos, E. (2008). SVD based Initialization: Ahead Start for Nonnegative Matrix Factorization. *Pattern Recognition*, 41:1350–1362.

[10] Bregonzio, M., Gong, S., and Xiang, T. (2009). Recognising Action as Clouds of

Space-Time Interest Points. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshop*, pages 1948–1955.

[11] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[12] Chen, Y., Bi, J., and Wang, J. Z. (2006). Miles: Multiple-Instance Learning via Embedded Instance Selection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947.

[13] Dalal, N. and Triggs, B. (2005). Histograms of Oriented Gradients for Human Detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 886–893.

[14] Dietterich, T., Lathrop, R., and Lozano-Pérez, T. (1997). Solving the Multiple Instance Problem with Axis-Parallel Rectangles. *Artificial Intelligence*, 89(1/2):31–71.

[15] Dollár, P., Rabaud, V., Cottrell, G., and Belongie, S. (2005). Behavior Recognition via Sparse Spatio-Temporal Features. In *Proc. Intern. Conf. on Computer Communications and Networks*, pages 65–72.

[16] Donoho, D. L. (2006). Compressed Sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306.

[17] Edelman, S. (1997). Computational Theories of Object Recognition. *Trends in Cognitive Sciences*, 1(8):296–304.

[18] Efros, A. A., Berg, A. C., Mori, G., and Malik, J. (2003). Recognizing Action at a Distance. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 726–734.

[19] Foulds, J. (2008). Learning Instance Weights in Multi-Instance Learning. Master's thesis, Department of Computer Science, University of Waikato.

[20] Foulds, J. and Frank, E. (2008). Revisiting Multiple-Instance Learning via Embedded Instance Selection. In *Proc Australasian Joint Conf. on Artificial Intelligence*.

[21] Gilbert, A., Illingworth, J., and Bowden, R. (2009). Fast Realistic Multi-Action Recognition using Mined Dense Spatio-Temporal Features. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 925–931.

[22] Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as Space-Time Shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

[23] Guillamet, D., Schiele, B., and Vitrià, J. (2002). Analyzing Non-Negative Matrix Factorization for Image Classification. In *Proc. Intern. Conf. on Pattern Recognition*, pages 116–119.

[24] Guillamet, D. and Vitrià, J. (2002). Classifying Faces with Non-Negative Matrix Factorization. In *Proc. Catalan Conf. on Artificial Intelligence*, pages 24–31.

[25] Heiler, M. and Schnörr, C. (2006). Learning Sparse Representations by Non-Negative Matrix Factorization and Sequential Cone Programming. *Journal of Machine Learning Research*, 7:1385–1407.

[26] Hoyer, P. O. (2004). Non-Negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, 5:1457–1469.

[27] Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A Biologically Inspired System for Action Recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*.

[28] Junejo, I. N., Dexter, E., Laptev, I., and Pérez, P. (2008). Cross-View Action Recognition from Temporal Self-Similarities. In *Proc. European Conf. on Computer Vision*, pages 293–306.

[29] Keeler, J., Rumelhart, D., and Leow, W. (1990). Integrated Segmentation and Recognition of Hand-Printed Numerals. In *Advances in Neural Information Processing Systems*, pages 557–563.

[30] Kläser, A., Marszałek, M., and Schmid, C. (2008). A spatio-temporal descriptor based on 3d-gradients. In *Proc. British Machine Vision Conf.*, pages 995–1004.

[31] Kovashka, A. and Grauman, K. (2010). Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2046–2053.

[32] Laptev, I. and Lindeberg, T. (2003). Space-Time Interest Points. In *Proc. IEEE Intern. Conf. on Computer Vision*, volume 1, pages 432–439.

[33] Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning Realistic Human Actions from Movies. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[34] Lee, D. D. and Seung, H. S. (1999). Learning the Parts of Objects by Non-Negative Matrix Factorization. *Nature*, 401(6755):788–791.

[35] Lee, D. D. and Seung, H. S. (2000). Algorithms for Non-Negative Matrix Factorization. In *Advances in Neural Information Processing Systems*, pages 556–562.

[36] Lin, Z., Hua, G., and Davis, L. (2009a). Multiple Instance Feature for Robust Part-Based Object Detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 405–412.

[37] Lin, Z., Jiang, Z., and Davis, L. S. (2009b). Recognizing Actions by Shape-Motion Prototype Trees. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 444–451.

[38] Liu, J., Ali, S., and Shah, M. (2008). Recognizing Human Actions using Multiple Features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[39] Maron, O. (1998). *Learning from Ambiguity*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

[40] Maron, O. and Lozano-Pérez, T. (1998). A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems*, volume 28, pages 570–579.

[41] Maron, O. and Ratan, A. (1998). Multiple-Instance Learning for Natural Scene Classification. In *Proc. Intern. Conf. on Machine Learning*, pages 341–349.

[42] Mauthner, T., Roth, P., and Bischof, H. (2010). Temporal Feature Weighting for Prototype-Based Action Recognition. In *Proc. Asian Conf. on Computer Vision*.

[43] Mauthner, T., Roth, P. M., and Bischof, H. (2009). Instant Action Recognition. In *Proc. Scandinavian Conf. on Image Analysis*.

[44] Mikolajczyk, K. and Schmid, C. (2004). Comparison of Affine-Invariant Local Detectors and Descriptors. In *Proc. European Signal Processing Conf.*

[45] Mikolajczyk, K. and Uemura, H. (2008). Action Recognition with Motion-Appearance Vocabulary Forest. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[46] Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A Survey of Advances in Vision-Based Human Motion Capture and Analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126.

[47] Niebles, J. C. and Fei-Fei, L. (2007). A Hierarchical Model of Shape and Appearance for Human Action Classification. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[48] Niebles, J. C., Wang, H., and Fei-Fei, L. (2008). Unsupervised Learning of Human Action Categories using Spatial-Temporal Words. *Intern. Journal of Computer Vision*, 79(3):299–318.

[49] Nowozin, S., Bakir, G., and Tsuda, K. (2007). Discriminative Subsequence Mining for Action Classification. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 1919–1923.

[50] Paatero, P. (1997). Least Squares Formulation of Robust Non-Negative Factor Analysis. *Chemometrics and Intelligent Labratory Systems*, 37:23–35.

[51] Poppe, R. (2007). Vision-Based Human Motion Analysis: An Overview. *Computer Vision and Image Understanding*, 108(1-2):4 – 18.

[52] Quinlan, J. R. (1993). *C4.5 Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[53] Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action MATCH: a Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[54] Roth, P. M., Leistner, C., Berger, A., and Bischof, H. (2010). Multiple Instance Learning from Multiple Cameras. In *Proc. IEEE Workshop on Camera Networks (CVPR)*.

[55] Roth, P. M., Mauthner, T., Khan, I., and Bischof, H. (2009). Efficient Human Action Recognition by Cascaded Linear Classification. In *Proc. IEEE Workshop on Video-Oriented Object and Event Classification*.

[56] Ryoo, M. S. and Aggarwal, J. K. (2009). Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In *Proc. IEEE Intern. Conf. on Computer Vision*.

[57] Samaria, F. and Harter, A. (1994). Parameterisation of a Stochastic Model for Human Face Identification. In *Proc. of IEEE Workshop on Applications of Computer Vision*, pages 138–142.

[58] Schindler, K. and van Gool, L. (2008). Action Snippets: How Many Frames does Human Action Recognition Require? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[59] Schüldt, C., Laptev, I., and Caputo, B. (2004). Recognizing Human Actions: A Local SVM Approach. In *Proc. Intern. Conf. on Pattern Recognition*.

[60] Shahnaz, F., Berry, M. W., Pauca, V. P., and Plemmons, R. J. (2006). Document Clustering using Nonnegative Matrix Factorization. *Journal of Information Processing and Management*, 42:373–386.

[61] Smeaton, A. F., Over, P., and Kraaij, W. (2009). High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Divakaran, A., editor, *Multimedia Content Analysis, Theory and Applications*, pages 151–174. Springer Verlag.

[62] Thurau, C. and Hlaváč, V. (2008). Pose Primitive Based Human Action Recognition in Videos or Still Images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*.

[63] Viola, P. and Jones, M. (2002). Robust Real-Time Object Detection. *Intern. Journal of Computer Vision*, 57(2):137–154.

[64] Viola, P., Platt, J. C., and Zhang, C. (2006). Multiple Instance Boosting for Object Detection. In *Advances in Neural Information Processing Systems*, pages 1417–1424.

[65] Wang, H., Ullah, M. M., Kläser, A., Laptev, I., and Schmid, C. (2009). Evaluation of Local Spatio-Temporal Features for Action Recognition. In *Proc. British Machine Vision Conf.*

[66] Wang, J. and Zucker, J.-D. (2000). Solving the Multiple-Instance Problem: A Lazy Learning Approach. In *Proc. Intern. Conf. on Machine Learning*, pages 1119–1125.

[67] Weber, M., Welling, M., and Perona, P. (2000). Unsupervised Learning of Models for Recognition. In *Proc. European Conf. on Computer Vision*, pages 18–32.

[68] Weidmann, N., Frank, E., and Pfahringer, B. (2003). A Two-Level Learning Method for Generalized Multi-Instance Problems. In *Proc. European Conf. on Machine Learning*, pages 468–479.

[69] Weinberger, K. Q., Blitzer, J., and Saul, L. K. (2005). Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *Advances in Neural Information Processing Systems*.

[70] Weinland, D., Ronfard, R., and Boyer, E. (2006). Free Viewpoint Action Recognition using Motion History Volumes. *Computer Vision and Image Understanding*, 104:249–257.

[71] Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., and Ma, Y. (2009). Robust Face Recognition via Sparse Representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31:210–227.

[72] Yao, A., Gall, J., and Gool, L. V. (2010). A Hough Transform-Based Voting Framework for Action Recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2061–2068.

[73] Yao, B. and Zhu, S.-C. (2009). Learning Deformable Action Templates from Cluttered Videos. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 1507–1514.

[74] Yeffet, L. and Wolf, L. (2009). Local Trinary Patterns for Human Action Recognition. In *Proc. IEEE Intern. Conf. on Computer Vision*, pages 492–497.

[75] Zach, C., Pock, T., and Bischof, H. (2007). A Duality Based Approach for Realtime TV-L1 Optical Flow. In *Proc. DAGM Symposium*.

[76] Zhang, L. and Zhou, W. (2010). On the Sparseness of 1-Norm Support Vector Machines. *Neural Network*, 23(3):373–385.

[77] Zhou, Z.-H., Sun, Y.-Y., and Li, Y.-F. (2009). Multi-Instance Learning by Treating Instances as Non-I.I.D. Samples. In *Proc. Intern. Conf. on Machine Learning*, pages 1249–1256.

[78] Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. (2003). 1-Norm Support Vector Machines. In *Advances in Neural Information Processing Systems*.

[79] Zou, H. (2007). An Improved 1-Norm SVM for Simultaneous Classification and Variable Selection. *Journal of Machine Learning Research*, 2:675–681.