RECOGNITION OF COMPLEX EVENTS IN OPEN-SOURCE WEB-SCALE VIDEOS: FEATURES, INTERMEDIATE REPRESENTATIONS AND THEIR TEMPORAL INTERACTIONS

by

SUBHABRATA BHATTACHARYA
B.E. Burdwan University (India), 2003

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Center for Research in Computer Vision
in the College of Electrical Engineering and Computer Science
at the University of Central Florida
Orlando, Florida

Summer Term
2013

Major Professor: Mubarak A. Shah

ABSTRACT

Recognition of complex events in consumer uploaded Internet videos, captured under real-world settings, has emerged as a challenging area of research across both computer vision and multimedia community. In this dissertation, we present a systematic decomposition of com-plex events into hierarchical components and make an in-depth analysis of how existing research are being used to cater to various levels of this hierarchy and identify three key stages where we make novel contributions, keeping complex events in focus. These are listed as follows: (a) Extraction of novel semi-global features – firstly, we introduce a Lie-algebra based representation of dominant camera motion present while capturing videos and show how this can be used as a complementary feature for video analysis. Secondly, we propose compact clip level descriptors of a video based on covariance of appearance and motion features which we further use in a sparse coding framework to recognize realistic actions and gestures. (b) Construction of intermediate representations – We propose an efficient probabilistic representation from low-level features computed from videos, based on Maximum Likelihood Estimates which demonstrates state of the art performance in large scale visual concept detection, and finally, (c) Modeling temporal interactions between intermediate concepts – Using block Hankel matrices and harmonic analysis of slowly evolving Linear Dynamical Systems, we propose two new discriminative feature spaces for complex event recognition and demonstrate significantly improved recognition rates over previously proposed approaches.

To my wife Bipasha and my son Swamantak,

for their love and continual support.

$\sim$

To my parents and sister for their sacrifices.

ACKNOWLEDGMENTS

I would like to acknowledge my sincere gratitude to my amazing advisor Prof. Mubarak Shah for giving me an opportunity to work in the Computer Vision Lab. I would like to thank him for taking me under his aegis, providing valuable guidance and supporting me. During the course of this thesis, I had the fortune to get introduced to Dr. Rahul Sukthankar, who has been my co-advisor, a selfless promoter of my work and a constant source of motivation, both in professional and personal front. I would like to thank him and Dr. Zehngyou Zhang who have given me invaluable opportunities to work in Intel Labs and Microsoft Research, respectively as summer intern, which helped me closely observe how industrial research is pursued. I would like to extend my regards to Prof. Ratan Guha, Dr. Brian Moore, and Dr. Joseph Laviola Jr. for serving as my PhD committee members and for their valuable suggestions.

I am fortunate to have collaborated with Dr. C. Mohan and Dr. Manish Gupta of IBM Research, and late Dr. Anirban Chakrabarti and Dr. Shubhashis Sengupta of Infosys Technologies Ltd., in India, before pursuing graduate study. The experience I gathered there helped me apply my programming skills to practical research problems. I also owe a lot to my undergraduate thesis advisor Prof. Bhabatosh Chanda (Indian Statistical Institute) who introduced me to the amazing field of research in computer vision. Among my friends and colleagues, I would personally like to thank Dr. Mukundan Venkataraman, Dr. Ramin Mehran, and Dr. Kishore Reddy who have provided numerous technical insights and incomparable patronage during my research.

I thank my family – Baba, Ma, Bipasha, Swamantak and Amrita for their unconditional love and continual inspiration . Last but not the least, I am grateful to our family friend Rangan for his support, especially during harsh times.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

The Internet has seen a sudden surge in consumer video traffic since last two decades, primarily due to the increasing convenience of sharing of multimedia data coupled with the plummeting cost of equipment required to capture them. According to statistics [4] released by YouTube - the most popular video sharing portal, $103,680$ hours of multimedia content is uploaded everyday. Most of the videos uploaded in YouTube, are captured by amateur users with limited cinematographic knowledge, and are subjected to camera motion, background clutter and frequent illumination changes. Usually these videos depict high-level social events - such as a music concert, birthday party or instructional events such as cooking a recipe or teaching a piano lesson. Thus, sifting through such an enormous collection of videos for a specific event is a crucial task and is often painstakingly frustrating given the technological maturity of current video browsing algorithms. Most algorithms, rely heavily on the generosity of the uploader to provide meaningful textual labels relevant to the uploaded video content. Since such textual labels are frequently noisy [152, 177], automatic analysis of such videos are gradually attracting a lot of researchers from computer vision and multimedia communities.

One task within the realm of automatic video content analysis is the recognition of complex events contained in the videos. The goal of complex event recognition is to automatically detect high-level events in a given video sequence. However, due to the fast growing popularity of such videos, especially on the Web, solutions to this problem are in high demands. A feasible solution can directly cater to the needs of several target applications. In addition to the obvious benefit of making video search and retrieval more efficient and rewarding experience for the user some of the other applications are enlisted as follows:

- **Product Promotion:** Tracking user interest based on the video content they watch - to promote advertisement of certain products.

- **Video Virality Prediction:** Helping broadcast agencies predict important statistics about a video such as virality of views, geographical location of viewers etc. moments after a video is uploaded – thereby optimizing broadcast channel bandwidth.

- **Textual Summarization:** Enabling human observers with meaningful textual recounting of a video in a relatively short time without substantial human intervention.

In this dissertation, we introduce a bottom-up approach towards decomposing the problem of complex event recognition into several intermediate steps, which can either be used independently or in the proposed coherent framework to improve the overall performance of the recognition. The problem of recognizing "Complex events" or "High-level events" in videos, as explained in [68], encloses a relatively more mature area of research – recognizing actions [32, 51, 66, 71, 73–75, 78, 84, 119–121, 141, 143, 190, 191, 194]. Although the terms – events, actions, interactions, activities, and behaviors have been used interchangeably in the literature [5, 24], there is no agreement on the precise definition of each term. We define high-level or complex events as long-term spatially and temporally dynamic object interactions that happen under certain scene settings [68]. Thus, in order to propose a feasible solution that caters to recognition of complex events, a hierarchical decomposition of an event is the foremost important step. The following paragraph and the accompanying Fig. 1.1, reiterates our views on this hierarchical structure of a complex event, on which the subsequent contributions of this dissertation are based.

At the lowest end of the hierarchy, we have movement : "an entity (e.g. hand) is moved with large displacement in right direction with slow speed". Movements can also be referred to as attributes which have been recently used in human action recognition [97] following their successful use in face recognition in a single image. Next are activities or actions, which are sequences of movements (e.g. "hand moving to right followed by hand moving to left", which is a "waving" action). An action has a more meaningful interpretation and is often performed by entities (e.g., human, animal, and vehicle). An action can also be performed between two or more

entities, which is commonly referred to as an *interaction* (e.g. person lifts an object, person kisses another person, car enters facility, etc.). Motion verbs can also be used to describe interactions. Recently the Mind's eye dataset is released under a DARPA program which contains many motion verbs such as "approach", "lift" etc [13]. In this hierarchy, *concepts* span across both actions and interactions. In general, *concept* is a loaded word, which has been used to represent objects, scenes, and events, such as those defined in LSCOM (Large-Scale Concept Ontology for Multimedia) [112]. Finally, at the top level of the hierarchy, we have *complex* or *high-level* events that have larger temporal durations and consist of a sequence of interactions or stand-alone actions, e.g., an event "changing a vehicle tire" contains a sequence of interactions such as "person opening trunk" and "person using wrench", followed by actions such as "squatting" and so on. Similarly, another complex event such as "birthday party" may involve actions like "person clapping" and "person singing", followed by interactions like "person blowing candle" and "person cutting cake". Note that although we have attempted to encapsulate most semantic components of complex events in a single hierarchy, because of the polysemous nature of the words, adopting the same terminologies in the research community is an impossible objective to achieve.

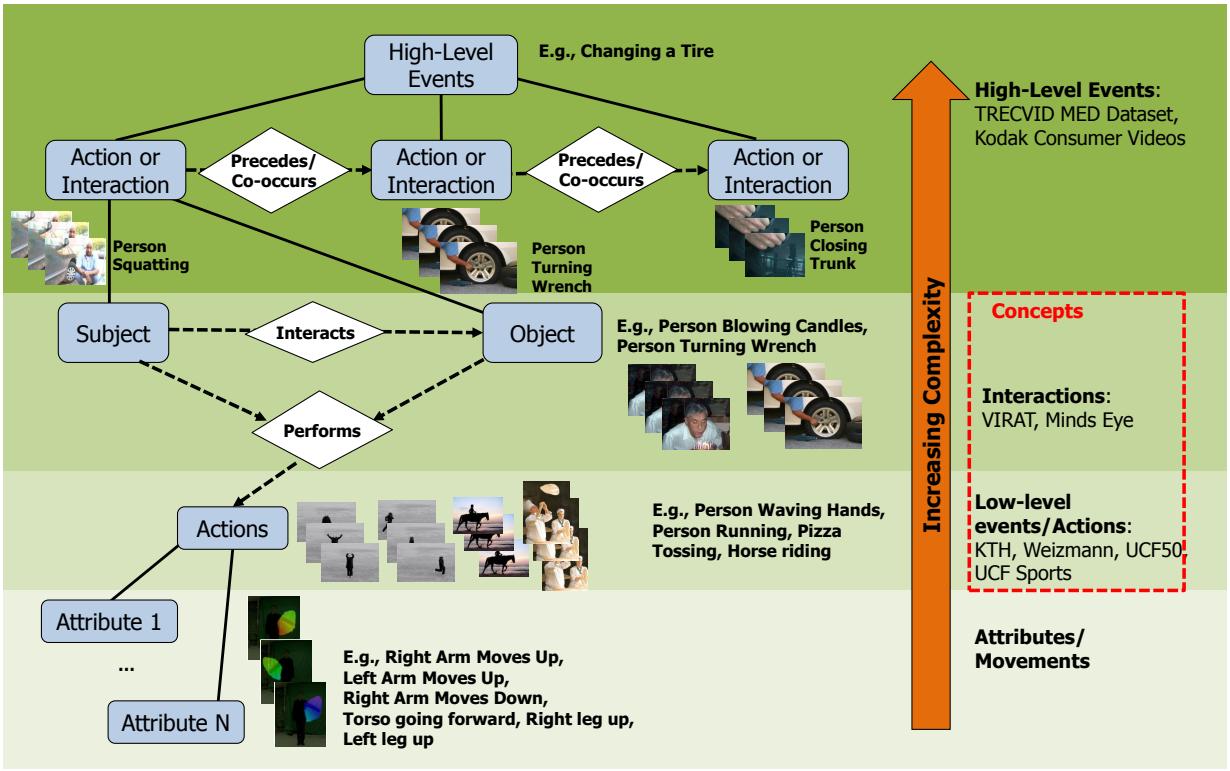


Figure 1.1: A taxonomy of semantic categories in videos, with increased complexity from bottom to top. *Attributes* are basic components (e.g., movements) of *actions*, while actions are key elements of *interactions*. *High-level events* (focus of this dissertation) lie on top of the hierarchy, which contain (normally multiple) complex actions and interactions evolving over time.

1.1 Technical Challenges

Current approaches heavily rely on classifier-based method employing directly computable low-level features from videos. Research strongly suggests the joint use of multiple features, such as static frame-based features [37, 101, 123], spatio-temporal features [39, 77, 81, 145, 170], and acoustic features [11, 85, 104, 127, 185]. However, with the popularity of handheld video recording devices, a huge amount of videos are currently being captured by amateur users under unconstrained conditions with limited quality control (in contrast to videos from broadcast news, documentary, or controlled surveillance, for example). Since the low-level features proposed in

earlier research are designed with more controlled conditions in mind [23, 144, 178], it is still not clearly understood [24, 34, 67, 113, 115] if they are comprehensive enough to capture discriminative information exhaustively from open-source videos.

Once features are computed, they are typically quantized into "video words" and each video is reduced to a histogram of video words also known as the bag-of-video-words (BoVW) representation. Once such a representation is obtained, classifiers are trained to establish a correspondence between the quantized layer on the features, to the actual label of the event depicted in the video. Although BoVW has shown promising retrieval results [36, 69], methods based on this paradigm suffer from the usual disadvantages of quantization used in converting raw features to discrete codewords as pointed out in [131, 162, 167]. Fig. 1.2 shows a generic framework for Bag-of-visual-words representation using different audio-visual features, predominantly used in various video analysis applications.

Another drawback of BoVW representations do not have any semantic understanding of the hierarchical components, such as interactions or actions that constitute the complex event. Needless to say, the sense of spatio-temporal localization of these components are lost in this coarse representation. Thus, this representation is unable to bridge the semantic gap between computable low-level features (e.g. visual, audio, and textual features) and semantic information that they encode (e.g. the presence of meaningful classes such as "a person clapping", "sound of a crowd cheering", etc.) [149]. This is why, with much progress made in the past decade in this context, the computational approaches involved in complex event recognition are reliable only under certain domain-specific constraints.

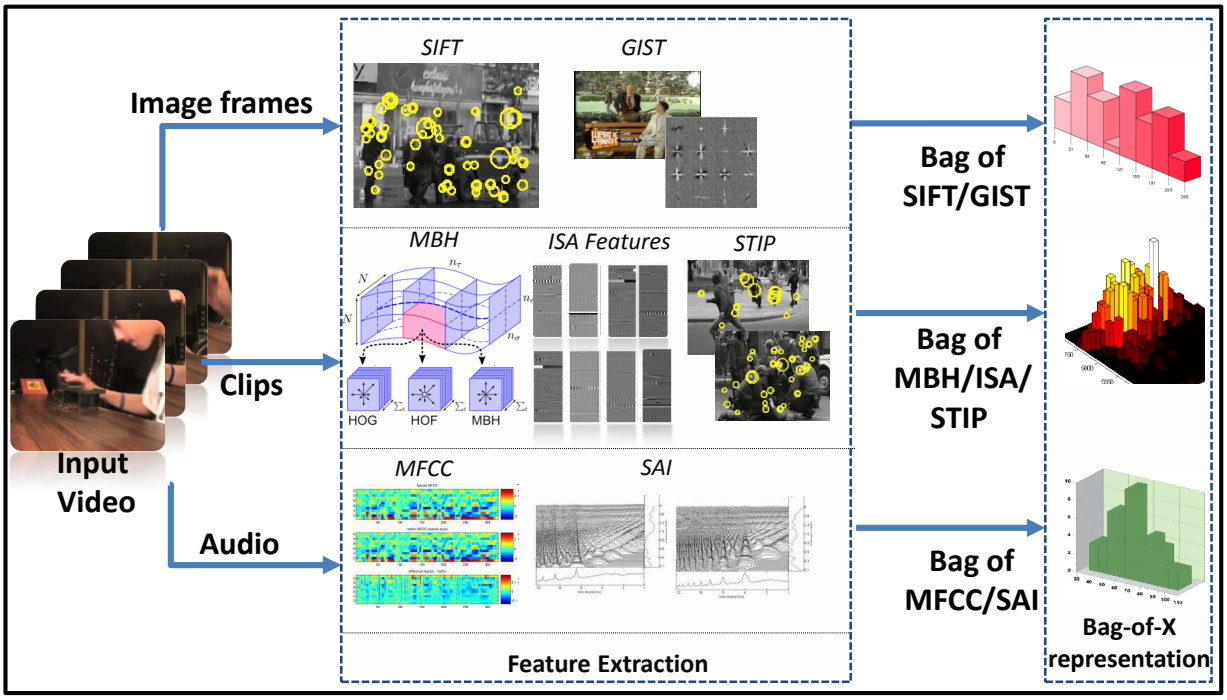


Figure 1.2: Bag-of-features representations obtained from different feature modalities for high-level event detection.

In order to overcome the problems encountered by BoVW methods based on purely low-level features, researchers have proposed the use of semantic concepts [34, 67, 113, 115], where each model in the first layer detects a semantic concept, and a second-level model is used to recognize event classes using a representation based on the first-layer outputs as feature. Such hierarchical intermediate representations facilitates the use of sophisticated probabilistic graphical models [35, 59, 60, 62, 92, 114, 172, 184, 186] as final event classifiers, which can also handle temporal and causal relationships between these intermediate concepts, thereby providing a more meaningful structure to the event model. Fig. 1.3 illustrates a computational framework using simple graphical models (discrete Hidden Markov Models [92, 114, 184, 187]) for complex event recognition. Although these models are mathematically intuitive, they are computationally complex and require extensive training coupled with substantial domain knowledge.

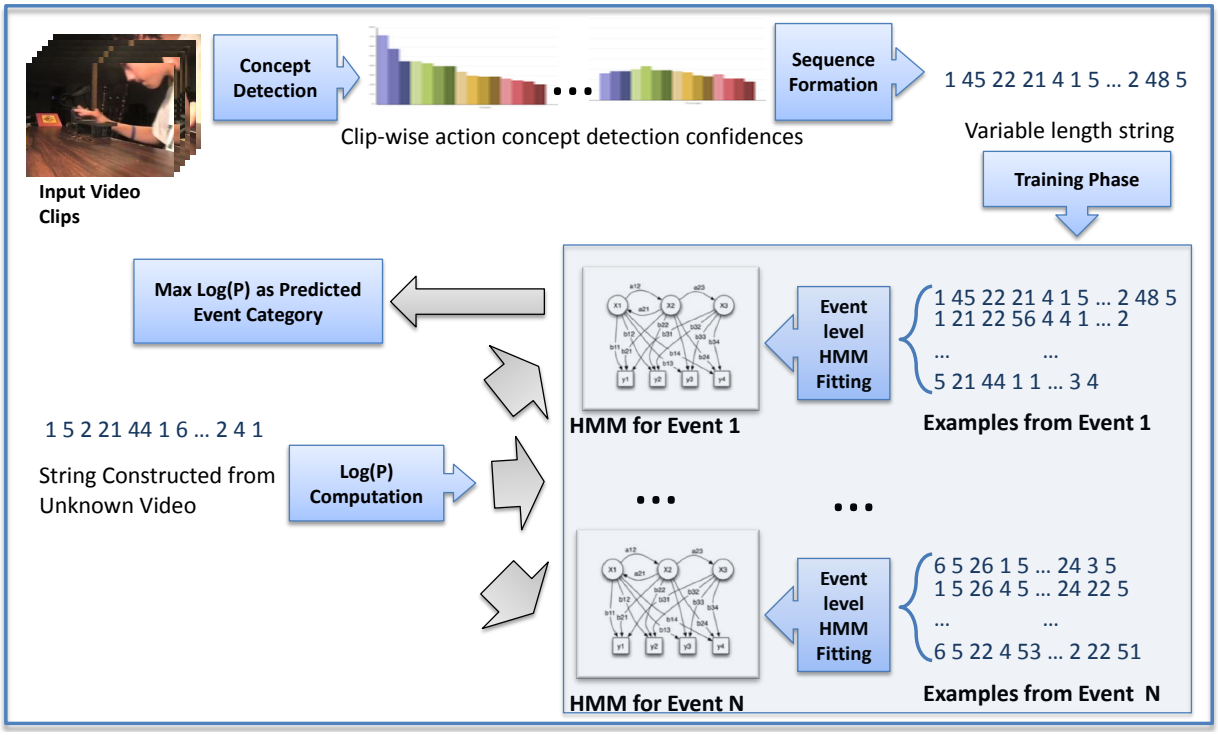


Figure 1.3: **A typical hierarchical classification computational approach using concepts and Hidden Markov Models (HMM):** An input video is divided into small overlapping clips on which, concept detectors are applied. The concept detector confidence scores are then discretized into a symbol sequence where each symbol denotes the presence of a concept detected with maximum confidence. The symbol sequence is input to different HMMs pretrained using symbol sequences generated from training data. The model that generates the maximum likelihood, given the input sequence identifies the true class of complex event depicted in the video.

1.2 Objectives

As elaborated in Section 1.1, recognition of complex events in unconstrained environments is far from being a solved problem in computer vision and involves substantial understanding at every level of the event hierarchy presented in Fig. 1.1. The goal of this dissertation is to explore these levels in detail to come up with novel methodologies that can be exploited to address the problem of complex event recognition in a computationally effective manner. Having said that, we summarize the key objectives we contrive to achieve, as the following :

- **Design of features** that capture relevant information from videos, and are robust to camera motion, illuminance changes, background clutter etc.

- **Engineer computationally efficient intermediate representations** on top of already existing low-level features, that can be easily integrated into semantically meaningful spatio-temporal concepts.

- **Formulate complex event models** based on temporal dynamics of the mid-level spatio-temporal concepts.

Based on our observations, in this dissertation we broach forward a systematic decomposition of complex events and make practically viable contributions at all levels of event representation hierarchy, to demonstrate promising improvement in final recognition performance.

The rest of this dissertation elucidates the claims made in the above statement starting with a summary of our key technical contributions.

1.3 Contributions

In this dissertation, we have addressed the problem of complex event recognition from unconstrained consumer videos and proposed solutions to different sub-problems at multiple tiers. To this end, we make some solid contributions, catering to a different area under the broader research area of complex event recognition. We align them to our objectives that are already listed in Section 1.2 as follows:

- **Design of features:** Within the purview of this effort, we explore two complementary sources of information to design features that are useful for content based video analysis in realistic scenarios. The first one is semi-global in nature, computed from small segments from the video, while the second one is based on ambient camera motion present during the video capture process.

For the semi-global clip-level descriptor [20], we compute kinematic features from optical flow and first and second-order derivatives of intensities to represent motion and appearance respectively. These low-level cues are then fused to construct covariance matrices which capture joint statistics between the distribution of motion and appearance of constituent pixels. Using an over-complete dictionary of the covariance based descriptors built from labeled training samples, we formulate human action recognition as a sparse linear approximation problem. Within this, we pose the sparse decomposition of a covariance matrix, which also conforms to the space of semi-positive definite matrices, as a determinant maximization problem. Also since covariance matrices lie on non-linear Riemannian manifolds, we compare our former approach with a sparse linear approximation alternative that is suitable for equivalent vector spaces of covariance matrices. This is done by searching for the best projection of the query data on a dictionary using an Orthogonal Matching pursuit algorithm.

For the camera motion based descriptor [19], we assume that a dominant homography exists between subsequent pairs of frames in a given video shot. Using purely image-based methods, we compute homography parameters that serve as coarse indicators of the camera motion. Next, using Lie algebra, we map the homography matrices to an intermediate vector space that preserves the intrinsic geometric structure of the transformation. Multiple time series are then constructed from these mappings. Features computed on these time series are used for discriminative classification of video shots. In addition, we provide an in-depth analysis of different features computed from time-series and their impact on the classification of different shots.

Our empirical evaluations of the proposed features on several challenging datasets demonstrate conclusive evidence towards their applicability in content based open-source video analysis applications.

- **Engineer computationally efficient intermediate representations:** We present an efficient

alternative [21] to the traditional vocabulary based on bag-of-visual words (BoVW) used for visual classification tasks. Our representation is both conceptually and computationally superior to the bag-of-visual words: (1) We iteratively generate a Maximum Likelihood estimate of an image given a set of characteristic features in contrast to the BoVW methods where an image is represented as a histogram of visual words, (2) We randomly sample a set of characteristic features called *anchors* instead of employing computation intensive clustering algorithms used during the vocabulary generation step of BoVW methods. Our performance compares favorably to the state-of-the-art on experiments over three challenging human action and a scene categorization dataset, demonstrating the universal applicability of our method. We also perform an extensive evaluation of the proposed method in context of spatio-temporal concept detection in YouTube videos and show definite improvement over traditional bag-of-feature based representations, popular in literature.

- **Formulate complex event models:** Here we represent each video depicting a complex event, as an ordered vector time-series, where each time-step is a vector containing confidences returned by a set of pre-trained spatio-temporal concept detectors. Using, foundations from linear dynamical systems, we extract two complementary features, the first is based on Block Hankel matrices, which captures dependencies between each observation vector, within the context of the entire time-series. The second exploits statistically meaningful characteristics from multiple interacting time-series such as lag-independence, harmonics, frequency proximity etc. Experiments conducted on NIST's, TRECVID datasets for Multimedia Event Detection (MED 2011 & MED 2012), demonstrate high-fidelity of our method [18] in modeling temporal interactions. In addition, our representation built on top of spatio-temporal concepts trained in a single feature modality, yield comparable results with the published state of the art in complex event recognition.

1.4 Organization

In order to make this dissertation comprehensive without compromising the legibility we have organized it in the following manner. In **Chapter 2** we provide a brief literature review on the topics discussed in Sections 1.1- 1.2, explicitly indicating the pertinence of the previous work to this context. The following chapters form the core technical discourse mentioned under contributions in Section 1.3. The design and implementation of good features for analysis of open-source videos is covered in**Chapters 3- 4**. This is followed by **Chapter 5** where we present a computationally efficient intermediate representation for spatio-temporal classification tasks. At the end of this chapter, we introduce a large collection of semantically interpretable spatio-temporal concepts that can leverage the proposed intermediate representation. In **Chapter 6**, we introduce a novel discriminative representation to model the temporal dynamics between spatio-temporal concepts, and set the context for complex event recognition in consumer videos. Finally, we provide some insights towards future work in **Chapter 7**, summarizing our observations in **Chapter 8**.

CHAPTER 2: LITERATURE REVIEW

In order to set the context for complex event recognition in consumer videos as a research problem, we first begin with the review of some of the prominent approaches. For brevity, we segregate this chapter into several sections, each highlighting prior research work in different directions that are relevant to several sub-problems which we have addressed within the purview of this dissertation.

2.1 Complex Event Recognition In Open-Source Videos

There have been several related papers that review research of video content recognition. Most of them focused on human action/activity analysis, e.g., [5] by Aggarwal and Ryoo, [132] by Poppe and [161] by Turaga *et al.*, where low-level features, representations, classification models, and datasets were comprehensively surveyed. While most human activity research was done on constrained videos with limited content (e.g., clean background and no camera motion), recent works have also shifted focuses to the analysis of realistic videos such as user-uploaded videos on the Internet, or broadcast and documentary videos.

In [151], Snoek and Worring surveyed approaches to multimodal video indexing, focusing on methods for detecting various semantic concepts consisting of mainly objects and scenes. They also discussed video retrieval techniques exploring concept-based indexing, where the main application data domains were broadcast news and documentary videos. Brezeale and Cook [27] surveyed text, video, and audio features for classifying videos into a predefined set of genres, e.g., "sports" or "comedy". Morsillo et al. [109] presented a brief review that focused on efficient and scalable methods for annotating Web videos at various levels including objects, scenes, actions and high-level events. Lavee et al. [82] reviewed event modeling methods, mostly in the context of simple human activity analysis. A review more related to this dissertation is the one by Ballan *et*

al. [12], which discussed features and models for detecting both simple actions and complex events in videos.

Figure 2.1: Examples of TRECVID MED 2010 and 2011 events. In 2011, in addition to 10 events used for official evaluation, TRECVID also defined 5 events for system development (e.g, parameter tuning).

Motivated by the need of analyzing complex events in Internet videos, the annual NIST TRECVID [148] activity defined a new task in 2010 called Multimedia Event Detection (MED). Each year a new (or an extended) dataset is created for cross-site system comparison. Table 2.1 summarizes the 2010 - 2012 editions of TRECVID MED datasets. The MED data consists of user-generated content from Internet video hosting sites, collected and annotated by the Linguistic

Data Consortium (LDC[1]). Fig. 2.1 gives an example for each event class. In MED 2010, only three events were defined, all of which are long-term procedures. The number of classes increased to 15 in the much larger MED 2011 dataset. Out of the 15 classes, 5 are only annotated on the training set for system development (e.g., feature design and parameter tuning), and the remaining 10 are used in the official evaluation. Besides several procedure events, there are also a few social activity events included in 2011, e.g., "wedding ceremony" and "birthday party". MED 2012, released in early 2012, consists of 2000 training videos from 10 new event categories such as: *Attempting a bike trick*, *Cleaning an appliance*, etc. The current editions of MED data only contain binary event annotations on video-level, and the MED task is focused only on video-level event classification.

Table 2.1: Overview of TRECVID MED 2010-2012 [1] datasets. The TRECVID videos are available upon participation of the benchmark evaluation. For all the three datasets, the positive videos are evenly distributed in the training and test sets.

Dataset	# training/test videos	# classes	# +ve videos/class	Avg. Length	Size
MED 2010	1,746/1,741	3	89	119 s	38 GB
MED 2011	13,115/32,061	15	253	114 s	559 GB
MED 2012	9,145/50,715	10	241	124 s	664 GB

Most of the successful complex event recognition approaches proposed till date [34,67,113, 115] advocate fusion of classifiers trained on bag-of-feature based representations from different information modalities. In [34,67], the authors introduced use of semantically interpretable spatial, spatio-temporal, and audio concepts in addition to bag-of-feature representations to achieve high event recognition accuracies. Recently, Inoue *et al.* [61] reported promising results in TRECVID MED task by using HMMs to characterize sequence of audio concepts.

In order to improve the performance of future approaches that strive to address this problem, we need to pay significant attention at several stages of the event recognition work-flow. Reiterating what we have already learned in Section 1.2, these are: (1) Design of features that capture

[1] http://www.ldc.upenn.edu/

relevant information from videos, and are robust to camera motion, illuminance changes, background clutter etc., (2) Engineering computationally efficient intermediate representations on top of already existing low-level features, (3) Integrating intermediate representations into mid-level spatio-temporal concepts that are semantically interpretable while modeling complex events, and (4) Formulate complex event models based on temporal dynamics of the mid-level spatio-temporal concepts.

Since feature extraction forms the first and foremost requirement in most video analysis applications, we elaborate on some of the predominantly used techniques from computer vision literature in the next section. In addition, we discuss how our proposed feature design addresses some of the problems that are not addressed by the state of the art.

2.2 Extraction of Relevant Features

Feature extraction is unarguably very crucial for event recognition as introduction of noise at the earliest stage of the recognition process can result in undesirable performance in the final classification. Research in action or event recognition has addressed this problem in different ways. Early efforts include [39, 81] where the authors introduce special detectors capable of capturing salient change in pixel intensity or gradients in a space-time video volume and later describing these special points or regions using statistics obtained from neighboring pixels. Direct extension of interest point based approaches from images such as 3D-SIFT [145](a space time adaptation of the SIFT [101] descriptor), HOG3D [77](a Spatio-Temporal Descriptor based on 3D Gradients derived from the principles of the HOG [37] descriptor for Human detection), Hessian STIP [180] (a Hessian extension of the SURF [15] key-point detector to incorporate temporal discriminativity); are some of the proposed alternatives. Recently, Weng and colleagues introduced motion boundary histograms [170] that exploits the motion information available from dense trajectories.

These interest point based approaches are incorporated into a traditional bag of video words

framework [155] to obtain an intermediate representation of a video that can further be used in a supervised [46] or un-supervised classification [58] algorithm for recognition purposes. While these approaches have been proved to be successful in context of event recognition, since they rely on highly localized statistics over a small spatio-temporal neighborhood e.g. $50 \times 50 \times 20$ [39, 170] relative to the whole video, different physical motions within this small aperture, are indistinguishable.

2.2.1 Covariance Matrix as Spatio-temporal Feature Descriptor

Covariance matrices as feature descriptors, have been used by computer vision researchers in the past in a wide variety of interesting areas such as: object detection [133, 163, 164, 192], face recognition [125, 146], object tracking [90, 134], etc. The authors of [163] introduced the idea of capturing low-level appearance based features from an image region into a covariance matrix which they used in a sophisticated template matching scheme to perform object detection. Inspired by the encouraging results, a license plate recognition algorithm is proposed in [133] based on a three-layer, 28-input feed-forward back propagation neural network. The idea of object detection is further refined into human detection in still images [164] and videos [192]. In [164], Tuzel *et al.* represented the space of d-dimensional nonsingular covariance matrices extracted from training human patches, as connected Riemannian manifold. A priori information about the geometry of manifold is integrated in a Logitboost algorithm to achieve impressive detection results on two challenging pedestrian datasets. This was later extended in [192] to perform detection of humans in videos, incorporating temporal information available from subsequent frames.

The authors of [125] used the idea of using region covariance matrices as descriptors for human faces, where features were computed from responses of Gabor filters of 40 different configurations. Later, Sivalingam *et al.* proposed an algorithm [146] based on sparse coding of covariance matrices extracted from human faces, at their original space without performing any exponential mapping as proposed in previous approaches [125, 133, 163, 164, 192]. In their ap-

proach, the authors formulated the sparse decomposition of positive definite matrices as convex optimization problems, which fall under the category of determinant maximization (MAXDET) problems. Although we are inspired by the formulation proposed in [146], our work attempts to solve a completely different problem.

In a different vein, Porikli and Tuzel [134] came up with another application of region covariance matrices in context of tracking detected objects in a video. In their technique, the authors capture the spatial and statistical properties as well as their correlation of different features in a compact model (covariance matrix). Finally, a model update scheme is proposed using the Lie group structure of the positive definite matrices which effectively adapts to the undergoing object deformations and appearance changes. Recently, Li and Sun [90] extended the tracking framework proposed in [134], by representing an object as a third order tensor, further generalizing the covariance matrix, which in turn has better capability to capture the intrinsic structure of the image data. This tensor is further flattened and transformed to a reduced dimension on which the covariance matrix is computed. For adapting to the appearance changes of the object across time, the authors present an efficient, incremental model update mechanism.

That said, in context of event recognition in constrained web videos, the exploitation of covariance matrices as feature is relatively inchoate. Some earlier advances are discussed here in this particular direction in order to set the pertinence of this work to the interested reader. Along these lines, the authors of [57] proposed a methodology for detection of fire using covariance of features extracted from intensities, spatial and temporal information obtained from flame regions. A linear SVM was used to classify between a non-flame and a flame region in a video. Researchers [55, 56] have also attempted to classify elementary human actions [144] using descriptors based on covariance matrices. In contrast, our work addresses a more diverse and complex problem. To summarize, we make the following contributions in this work: (1) We propose a novel descriptor for video analysis which captures spatial and temporal variations coherently, (2) Our descriptor is flexible to be used for different application domains (unconstrained event recognition, gesture

recognition etc.), and (3) We propose two different classification strategies based on concepts from sparse representation that can be used in the recognition pipeline independently.

2.2.2 *Shot-level Camera Motion Descriptor based on Cinematographic Principles*

Researchers have explored the direction of analyzing video content based on camera motion in the past [136, 137]. In one of the earliest efforts [153], the authors qualitatively estimate camera pan, tilt, zoom, and roll from a sequence of images. [195] extends the idea to shots with camera rotation, where mutual information between motion vectors is utilized. In [126], Park *et al.* explored further using linear combination of motion vectors. While these techniques relied on optical flow to obtain motion vectors, a few teams in TRECVID 2005 [148] used motion vectors provided in MPEG stream for this purpose.

From a different perspective, Fablet *et al.* [44] make use of local spatio-temporal derivatives to classify dynamic content of shots without motion segmentation. Wang and Cheong on the other hand, explore the possibilities of using a Markov Random Field based motion foreground vs background labeling framework [171] together with cinematographic principles to classify pan, tilt, zoom, track and establishing shots. Approaches proposed in [2, 156, 173] focus on specific semantic classes of videos. For example in [173] the authors employ structure tensor histograms to determine motion characteristics in shots from action movies. Similarly, [2, 156] leveraged on specific cinematographic techniques that only applied to sports videos to address the shot classification problem.

In contrast to the previous approaches, our technique to represent camera motion has the following contributions: (1) We obtain global camera motion by robustly estimating frame to frame homographies unlike approaches [44, 126, 153, 195] that rely on local optical flow based techniques, which are often noisy or full structure from motion based approach [193],which is computationally expensive (2) Compared to approaches [156] that use homographies directly for classification, our lie-algebra based representation homographies is more accurate and computationally less expen-

sive, (3) Our global features computed from a shot consider temporal continuity between frames, are superior to orderless bag of words techniques used in [148], thereby eliminating any need for explicit temporal alignment of shots of unequal lengths, (4) Our representation is capable of classifying a broader category of shots as compared to [118, 126, 153, 193, 195]. As part of this work, we also introduce a dataset that consists of eight cinematographic shot classes [9] which is freely available to the research community, (5) Our method is more versatile than approaches suggested in [2, 156, 173] which apply to specific domains such as movies or sports. It also requires fewer parameters to adjust as compared to [171], which require explicit motion segmentation, and (6) Finally, this is the first work to show how camera motion alone could be used to detect high level events in a video without any knowledge of the content.

Once meaningful features are extracted, it is often desired to transform the features to a common space that is robust to outliers. This common space is formally known as the intermediate representation which is usually achieved by vector quantization techniques. This yields the popular bag-of-visual-words representation of an image or a video. There have been several innovations in the traditional bag-of-words model that have been used for several visual classification tasks. These advances could be categorized broadly into two levels: representation and classification. In the next section, we briefly provide an overview of the techniques broadly used in literature and propose an efficient alternative to the traditional bag-of-words approach.

2.3 Probabilistic Intermediate Representation

At the representation level, Jurie and Triggs [72] show that the clustering process (usually k-means) required during vocabulary generation, is only capable of encoding regions rich in descriptor space. They introduce a radius-based clustering that is capable of generating better codebooks for general scenes. The authors of [182] propose an algorithm for learning a compact visual vocabulary through an iterative pair-wise merging approach, resulting in visual words de-

scribed by Gaussian Mixture Models (GMMs). GMMs are also employed in the construction of adaptive class-specific vocabularies by Perronnin *et al.* [130] and Farquahar *et al.* [47].

Using hidden topic learning models, Bosch *et al.* [26] introduce a novel vocabulary construction technique that represents each image with a topic distribution vector. Inspired by the success of generative techniques like pLSA in [26], Perronnin *et al.* apply Fisher kernels [129] to image categorization. Furthermore in [98], the authors introduce a method based on maximization of mutual information to group semantically similar visual words resulting in an efficient vocabulary. Tuytelaars and Schmid [162] discretize the high-dimensional space of image features using an optimal lattice structure to create a compact bag of visual words representation for images.

At the classification level, Grauman and Darrell [54] present a pyramid match kernel function that maps unordered feature sets into a higher-dimensional space of multi-resolution histograms, projecting the classification problem into a weighted histogram intersection in that space. This approach is further adapted by Lazebnik *et al.* [83] to spatial pyramid features that preserve a rough spatial information within the codeword, which is beneficial for classification using a histogram intersection kernel.

Methods such as [96, 108] are also popular, where the classification stage is not independent on the representation stage. [108] uses an ensemble of randomized trees for codebook generation as opposed to expensive clustering, followed by employing a tree-based classifier for the recognition task.

One of the problems with the codebook approach, analyzed by [25, 167], is the hard assignment of cluster centers to the visual words in an image which is performed while generating the vocabulary. To this extent Van Gemert *et al.* [167] propose a method to model ambiguity in assigning codewords to images, thereby improving classification accuracy for natural images that have large variation in appearance. In our approach, we circumvent this problem by creating a representation that maximizes the likelihood of generating the visual words corresponding to an image using a kernel density estimator. In fact, Van Gemert *et al.*'s soft-assignment representa-

tion becomes a special case of our proposed method, where our representative visual words are set to the cluster centers from a pre-defined codebook and the algorithm terminated after a single iteration.

Our approach also bears some philosophical resemblance to [47], wherein the authors first associate GMMs with each visual word, whose parameters are iteratively tuned using an expectation maximization algorithm. However, their approach suffers from over-fitting, for which they need to apply additional regularization techniques. Our approach (as we show in the following sections) has fewer parameters, is guaranteed to converge to a global optimum, is not prone to over-fitting and does not require explicit regularization. Secondly, in contrast to [72, 83, 98, 167], our representation does not require an expensive clustering mechanism for codebook generation. While our proposed method can certainly utilize any existing codebook, we show that the anchors in our maximum likelihood representation can also be simply initialized on a randomly-sampled unique set of visual words from a given dataset.

Although the above discussed technique eliminates the disadvantageous of traditional bag of visual words based techniques, the intermediate representation that we achieve after it is applied, is not semantically interpretable. Earlier research [34, 67, 115] show definite boost in event classification performance when naive intermediate representations are augmented with an additional semantic layer. Furthermore, it is often desired to get a structured semantic understanding of the visually observable content of the videos. This motivates us to explore the idea of atomic spatio-temporal (action) concepts that can be reliably detected in videos, which provide the additional semantic layer of information.

2.4 Spatio-Temporal Concepts For Complex Event Recognition

Action or Low-level event recognition has been an active field of research in computer vision for the last few years. The interested reader is referred to [161] for a literature survey.

There are several interesting works [48, 102] demonstrating extremely high recognition accuracies over respective benchmark datasets [23, 144]. While significant performance increase on these benchmark datasets is an important direction that drives event recognition research, it is hard to derive conclusive evidence about the viability of the methods tested on these datasets on realistic recognition scenarios. Liu *et al.* [100] presented the YouTube-11 human action dataset which partially addressed this issue, by introducing video samples from consumer uploaded videos that contained cluttered background, jittery camera motion, realistic actions performed by non-actors. This work was superseded by the Hollywood human actions and scenes dataset [106] which contained 12 classes of edited, stabilized Hollywood movie clips depicting realistic human actions. Both datasets presented in [100, 106] have relatively more complicated human actions and intra-class variance compared to [23, 144]. In the same vein of [100, 106], recently, two more datasets UCF50, and HMDB51 [80] are introduced that contain 50 and 51 human action categories. Both UCF50 and [80] derive their video clip sources from YouTube (unedited) and movies/consumer videos (edited/ unedited) respectively. However, in all these datasets, action or low-level event recognition is treated in complete isolation without taking the spatio-temporal occurrence of different low-level events. Studies [67] indicate that this could often be a limiting factor for complex event recognition tasks.

In view of the above, some recent datasets [122],TRECVIDMED11 have been introduced to motivate research in complex event recognition. The authors of [122] broach a dataset on human actions with 23 annotated categories captured from videos replicating surveillance scenarios with very limited camera motion. TRECVID MED11 event corpus, on the other hand, is a relatively more challenging dataset in this respect. It has a collection of $2,062$ high-quality consumer videos depicting 15 complex audio-visual events such as "Feeding an animal" or "working on a sewing project", which have large amount of visual variance within same semantic class.

Event recognition in such complex videos is relatively a new field. Research in this direction includes use of purely low-level features [69] in a bag of visual words framework to ap-

proaches [43, 177] that involve a combination of low-level features and noisy human annotated text tags that are usually available in YouTube videos. Recently, the authors of [67] introduced the usage of low-level events (person walking, person running etc.) for complex event recognition (batting a baseball run). The role of these low-level or atomic events is shown to be extremely useful for semantic analysis of videos in [70, 112] using the Large Scale Concept Ontology in Multimedia(LSCOM) dataset. The LSCOM dataset is constructed from still keyframes from News and Broadcast videos and consists of slightly over 400 annotated concepts depicting the presence of objects (aircraft carrier, helicopter hovering, etc.), natural scenes(desert, mountain etc.), calamities (flood, tornado etc.), well-known people (George Bush, Yaseer Arafat etc.) and so on in video frames under consideration. While these concepts are exhaustive, they are based on single images. This limits the exploitation of information from temporal and audio modalities which are available across subsequent frames in a video clip. This is found [67] to be detrimental to the performance of detectors trained on these single image based concepts.

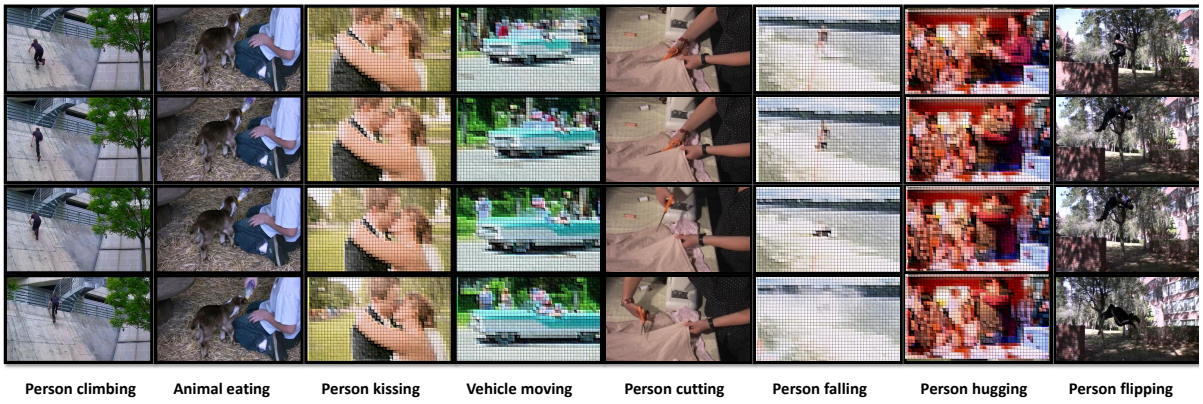


Figure 2.2: **Sample frames from the proposed low-level events dataset**: 8 different low-level event categories are shown here. Frames with human faces are intentionally pixellated to preserve privacy.

In this dissertation, we present a dataset consisting of a large number of generic low-level events (not necessarily all performed by human subjects), containing over $10,000$ exemplars. Some

sampled frames pertaining to a small subset of these 104 concepts are shown in Fig. 2.2 (actual video samples are shown in the accompanying supplementary material). This is followed by an in depth study of detectors trained using multiple information modalities. Finally, we demonstrate a practical application of these concept detectors on complex event recognition using two different approaches. In the first, we formulate event detection as an indexing problem in the high-dimensional space of concept detector confidences using feature trees. Whereas, in the second, we use hidden Markov models that exploit temporal relationship between low-level event detectors to predict a complex event category. We disseminate interesting insights through a large number of experiments which would be very useful for the research community.

Table 2.2: **Comparison with similar existing datasets:** Different high level characteristics of our low-level events dataset as compared to similar existing datasets. Our dataset outnumbers the existing ones by a significant margin, both in terms of number of examples $(10,000+)$ and number of classes $(100+)$.In addition, each video clip in this dataset is semantically related to a complex event from the original TRECVID MED 2011 event corpus, unlike clips from other datasets.

	Datasets				
Properties	**HOHA [106]**	**UCF50**	**VIRAT [122]**	**HMDB51 [80]**	**This Dataset**
Year	2009	2011	2011	2011	2012
num. Classes	12	50	12	51	104
num. Samples	3669	6400+	329	6766	$10,243$
Avg. Samples/class	142	100	167	100	95
Resolution	540×240	640×480	1920×1080	320×240	640×480
Action Type	Human	Human	Human	Human	Human(84)+ Misc.
Camera Motion	Smooth	Jitter	Aerial Jitter	Compensated	Jitter + Misc.

Our proposed low-level event dataset is designed after making careful consideration of the issues presented in [23, 80, 100, 106, 122, 144]. In contrast to the above, the low-level events in our dataset are selected from complex events in a completely top-down manner that leverages the fact that actions do not occur in isolation. In other words, we assume that atomic low-level events carry representative information of underlying events, thereby attempting to narrow the semantic divide between action and event recognition. We present some comparative statistics of our dataset with other existing datasets in Table 2.2. Our contributions in this work can be summarized as follows

: (1) We introduce a dataset of low-level events that is not only the largest in terms of number of categories and size, it also reflects real-world complex events, (2) We make a comprehensive study of existing approaches that can be leveraged to build reliable detectors trained on these low-level events and present a baseline method which can be used by action recognition researchers, (3) We elucidate some of the common practices that can be used to recognize complex events using the responses of our pre-trained low-level event detectors. To the best of our knowledge, this is the first work that motivates research in realistic low-level event recognition from the perspective of complex events, which has an important philosophical standpoint.

With such a rich spatio-temporal concept based semantic representation of a video, one can explore a multitude of research directions. Since, these concept detectors can applied on a video at smaller temporal granularities, we can observe the distribution of concepts at every measured temporal interval. This can in turn help obtain some useful statistics based on the various temporal interactions concepts have with each other. The statistics can later be used to model complex events. The next section closes the loop on complex event recognition by achieving a novel discriminative model that exploits the temporal dynamics between concepts using theoretical foundations from Linear Dynamical Systems.

2.5 Temporal Dynamics Of Spatio-Temporal Concepts

Temporal interactions between concepts have been represented using graphical models (directed and undirected) in the past extensively by researchers using Hidden Markov Models [92, 184, 187], Bayesian Networks (BNs) [59, 62], Conditional Random Fields (CRF) [35, 172], Dynamic Bayesian Networks (DBNs) [60] etc.

Over the past two decades, several other works [92, 114, 154, 184] have used HMMs and their variants in human action recognition. Starner and Pentland were among the early adopters of HMMs in their research [154] on sign language recognition. Xie et al. [184] demonstrated

how HMMs and hierarchical composition of multiple levels of HMMs could be efficiently used to classify play and non-play segments of soccer videos. Motivated by the success of HMMs, Li et al. [92] introduced an interesting methodology to model an action where hidden states in HMMs were replaced by visualizable salient poses (which forms an action) estimated using Gaussian mixture models. Since states in HMMs are not directly observable, mapping them to poses is an interesting idea. In the work by Natarajan and Nevatia [114], an action is modeled by a top-down approach, where the topmost level represents composite actions containing a single Markov chain, and the middle level represents primitive actions modeled using a variable transition HMM, followed by simple HMMs that form the bottommost layer representing human pose transitions. Recently, Inoue et al. [61] reported promising results in TRECVID MED task [1] by using HMMs to characterize audio which is often observed to be a useful cue in multimedia analysis.

There are other types of directed graphical models that have been studied in event recognition. Another disadvantage with the HMM formulation is its incapability to model causality. This problem is alleviated by a different kind of directed graphical model called Bayesian Networks (BN). BNs are capable of efficiently modeling causality using conditional independence between states. This methodology facilitates semantically and computationally efficient factorization of observation state space. In this vein, Intille and Bobick introduced an agent based probabilistic framework that exploits the temporal structure of complex activities typically depicted in American football plays [62]. They used noisy trajectory data from soccer players collected from a static overhead camera to obtain temporal (e.g., before or after) and logical (e.g., pass or no pass) relationships, which are then used to model interactions between multiple agents. Finally, the BNs are applied to identify 10 types of strategic plays.

BNs cannot implicitly encapsulate temporal information between different nodes or states in the finite state machine model. Dynamic Bayesian Networks (DBNs) can achieve this by exploiting the factorization principles available in Bayesian methods while preserving the temporal structure. Research on event recognition using DBNs is relatively new as compared to other ap-

proaches since it requires a certain amount of domain knowledge. In [60], Huang et al. presented a framework for semantic analysis of soccer videos using DBNs, where they successfully recognized events such as corner kicks, goals, penalty kicks, etc.

BNs, HMMs and their variants fall under the philosophy of generative classification, which models the input, reducing variance of parameter estimation at the expense of possibly introducing model bias. Because of the generative nature of the model, a distribution is learned over the possible observations given the state. However, during inference or classification, it is the observation that is provided. Hence, it is more intuitive to condition on the observation, rather than the state.

This has motivated researchers to investigate alternative strategies for modeling complex events using undirected graphical models, some of which are naturally suited for discriminative modeling tasks. To this end, Vail et al. [165] made a strong contribution by introducing Conditional Random Fields for activity recognition. In their work, the authors show that CRFs can be discriminatingly trained based on conditioning on the entire observation sequence rather than individually observed sample. A CRF can be perceived as a linear chain HMM without any directional edges between the hidden states and observations. In case of HMMs, the model parameters (transition, emission probabilities) are learned by maximizing the joint probability distribution, whereas, the parameters of a CRF (potentials) are learned by maximizing the conditional probability distribution. As a consequence, while learning the parameters of a CRF, modeling the distribution of the observations is not taken under consideration. The authors of [165] produced convincing evidence in favor of CRFs against HMMs in context of activity recognition. Inspired by the success of [165], Wang and Suter [172] introduced a variant of CRFs which can efficiently model the interactions between temporal order of human silhouette observations for complex event recognition. Wang and Mori [176] extended the idea of general CRFs to a max-margin hidden CRF for classification of human actions, where they model a human action as a global root template and a constellation of several "parts". More recently, in [35], Conolly proposed modeling and recognition of complex events using CRF, by taking observations obtained from multiple stereo systems under surveillance

domain.

Although undirected graphical models (CRFs) are far less complex than their directed counterparts (DBNs), and avail all the benefits of discriminative classification techniques, they are disadvantageous in situations where the dependency between an event/action and its predecessors or successors (e.g., cause and effect) needs to be modeled. Although some variants of CRFs can overcome this problem by incorporating additional constraints and complex parameter learning techniques, they are computationally slow.

While these models are mathematically elegant, most of them need extensive domain specific knowledge in addition to a large number of training samples. Also, since these models present an abstract layer over underlying intermediate representations (bag-of-features, concepts etc.), they are unable to handle imperfections present in the lower level, which is predominant in unconstrained scenarios. This motivates us to explore a different direction which is popular in other areas of computer vision [138, 160], yet has not received significant attention in context of complex event recognition. Our work emphasizes on extracting joint temporal evolution of underlying models which can be used in the recognition of complex events.

2.6 Summary

We reviewed some of the important topics necessary for understanding the problem of complex event recognition in this chapter. We began with an introduction to various low-level feature extraction techniques that are crucial for any high-level visual recognition tasks. We provided a brief overview on how our proposed feature extraction techniques can alleviate the problems faced by earlier research in this direction. Next, we provided an outline of methods that are used to construct a compact intermediate representation on top of the underlying features. We indicated some of the limitations of current approaches and introduced our alternative strategy which achieves this object. Thereafter, we elucidated semantically interpretable spatio-temporal concepts which

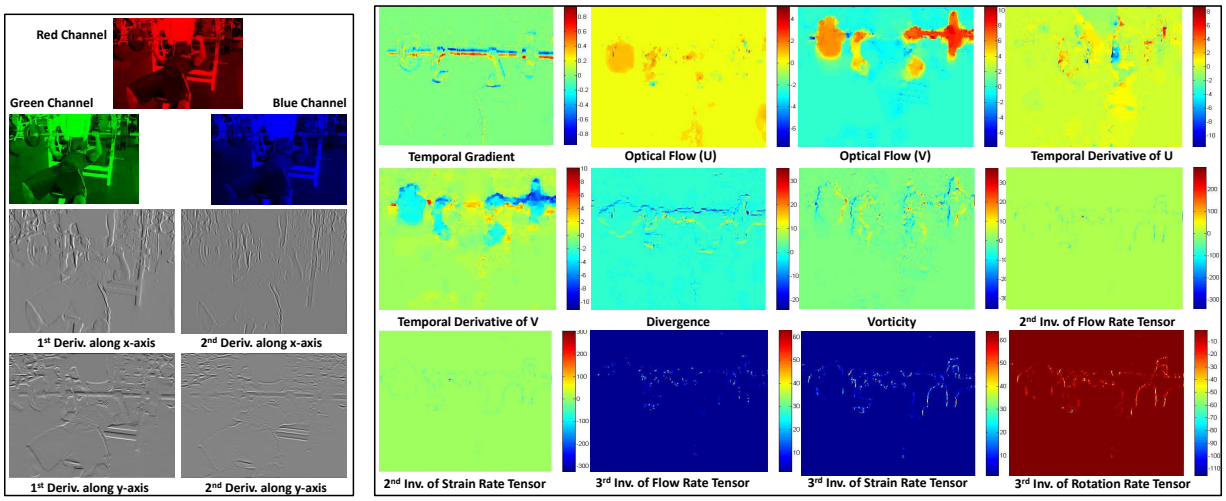
augment the capabilities of current intermediate representations in describing videos. Finally, we offered a literary sketch of existing techniques that exploit relationships between spatio-temporal concepts and argued how our novel representation can leverage this aspect to produce more discriminative complex event models.

CHAPTER 3: COVARIANCE OF MOTION AND APPEARANCE FEATURES FOR HUMAN ACTION AND GESTURE RECOGNITION

3.1 Introduction

Spatio-temporal feature extraction [39, 81, 170] is an extremely important stage in video analysis applications. Most of the prominent research split this stage into two equally important sub-stages namely, *Detection* and *Description*. The former refers to the search of interesting areas in a video that high degrees of discriminativity in terms of spatio-temporal information content. The latter refers to the quantification of the identified areas using common statistical tools. While describing the statistics of these small neighborhoods, often the temporal and the spatial information are treated independently. For e.g. the HOG-HOF descriptor used in [81] is generated by concatenating two independent histograms : the HOG (Histogram of Oriented Gradients [37])-contributing to the appearance (spatial) and the HOF (Histogram of Optical Flow) - contributing to motion (temporal). Doing so, the joint statistics between appearance and motion is lost which may be informative, particularly in case of action recognition in the practical scenarios where such information can be very useful. For e.g. consider the example of "pizza-tossing event" from the UCF50 [3] unconstrained actions dataset. Here, a circular white object undergoes a vertical motion which is discriminative for this event class. Precisely, the correlation between white object as captured by appearance features and its associated vertical motion captured basic and kinematic features is well explained in the covariance matrix than a concatenated 1-D histogram of the individual features. It is also important to note that contextual information available in the form of color, gradients etc., is often discriminative for certain action categories. Descriptors that are extensively gradient based such as HOG or HOF, need to be augmented with additional histograms such as color histograms to capture this discriminative information. Fig. 3.1 emphasizes how low-level features can be computed from different information modalities. Here, Figs. 3.1(a)

and 3.1(b) visualize the appearance and motion features respectively for a sample frame from the UCF50 dataset, where a person is exercising a "bench-press".



(a) (b)

Figure 3.1: **Low-level feature extraction from video clips:**(a) Appearance features,and (b) Motion features (basic and kinematic). The kinematic features are derived from optical flow and capture interesting aspect of motion with respect to a spatial neighborhood.

In view of the above, we propose a novel descriptor [20] for video event recognition which has the following properties: (1) Our descriptor is a concise representation of a temporal window/clip of subsequent frames from a video rather than localized spatio-temporal patches, for this reason, we do not need any specialized detectors as required by [39, 81, 180](2) It is based on an effective fusion of motion features such as optical flow and their derivatives, vorticity, divergence etc., and appearance feature such as first and second order derivatives of pixel intensities, which are complementary to each other. For this reason, descriptor can also be augmented to capture other complementary information available in videos e.g. audio, camera motion, (3) As the descriptor is based on joint distribution of samples from a set of contiguous frames without any spatial subsampling, it is implicitly robust to noise resulting due to slight changes in illumination, orientation etc. (4) It is capable of capturing the correlation between appearance with respect to motion and

vice-versa in contrast to concatenated 1-D histograms as proposed in [39, 77, 81, 145], also, since our final descriptor is based on the eigenvectors of the covariance matrix, they automatically transform our random vector of samples into statistically uncorrelated random variables, and (5) Finally being compact, fewer descriptors are required to represent a video compared to local descriptors and they need not be quantized.

It is the non-local and non-linear nature of our descriptor, that encourages us to explore intermediate representation and classification strategies other than the traditional vector quantization based bag-of-visual-words representation followed by SVM classification. In this context, we propose two sparse representation based techniques to perform low-level event recognition using these covariance matrices as atoms of an overcomplete dictionary. In the first one, we map the covariance matrices to an equivalent vector space using concepts from Riemannian manifold before building the dictionary. The classification is performed using a modified implementation of Orthogonal Matching Pursuit [159] which is specifically optimized for sparse-coding large sets of signals over the same dictionary. We compare this approach with a tensor sparse coding framework [146] formulated as a determinant maximization problem, which intrinsically maps these matrices to an exponential family. Although, our work is largely inspired by [163] and [146] in object recognition, to the best of our knowledge, ours is the first work that addresses low-level event recognition using a sparse coding framework based on covariance of motion and appearance features.

In order to make the chapter self contained, we briefly describe the theoretical details of all the phases involved in our low-level event recognition computation pipeline, beginning with the feature extraction step. Fig. 3.2 provides a schematic description of our approach [20] showing the steps involved in training phase (dashed blue box) and the testing phase (dashed red box).

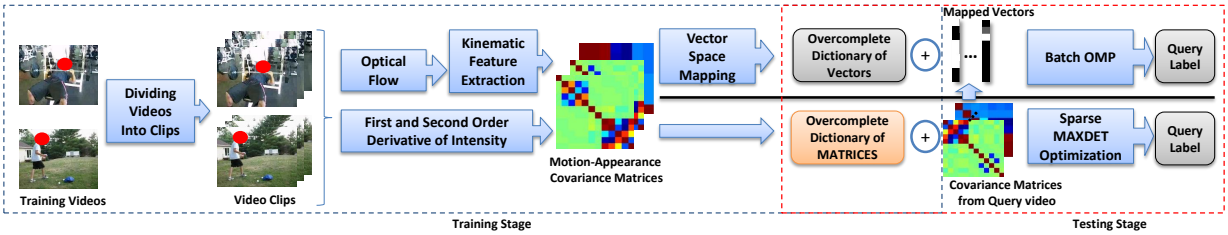


Figure 3.2: **An Overview of our approach [20]:** We begin with dividing training videos into multiple non-overlapping clips. Each clip is represented by a single covariance matrix computed from appearance and motion features as explained in Sections 3.2 and 3.3. A dictionary is created by stacking up the covariance matrices. Given, a test covariance matrix, its corresponding label is determined by solving a matrix determinant maximization problem as shown in Section 3.4.2. The final label for a video is obtained by aggregating the class labels predicted for individual clips.

3.2 Computation of Low-level Appeance and Motion Cues

Since our primary focus is on low-level event recognition in unconstrained scenarios, we attempt to exploit features from both appearance and motion modalities which provide vital cues about the nature of the event. Also since this chapter attempts to study how the appearance and motion change with respect to each other, it is important to extract features that are discriminative within a modality. Given a video, we split it into an ensemble of non-overlapping clips of N frames. For every pixel in each frame, we extract the normalized intensities in each channel, first and second order derivatives along the x and y axes. Thus every pixel at (x, y, t) can be expressed in the following vector form with $\mathbf{f_i}$, $\mathbf{f_g}$ denoting the intensity and its gradient components along the horizontal and vertical axes respectively, as:

$$\begin{aligned} \mathbf{f_i} &= \begin{bmatrix} R & G & B \end{bmatrix}^T, \\ \mathbf{f_g} &= \begin{bmatrix} \frac{\partial I}{\partial x} & \frac{\partial I}{\partial y} & \frac{\partial^2 I}{\partial x^2} & \frac{\partial^2 I}{\partial y^2} \end{bmatrix}^T, \end{aligned} \tag{3.1}$$

where R, G, B are the red, green, blue intensity channels and I being the gray scale equivalent of a particular frame. Fig. 3.1(a) visualizes the different low-level features that contribute to appearance description from a sample action frame from UCF50. Fig. 3.3 provides a 3-dimensional

visual illustration of covariance descriptors constructed using only the aforementioned appearance cues on 8 different action classes in UCF50 and their equivalent vector space representations.

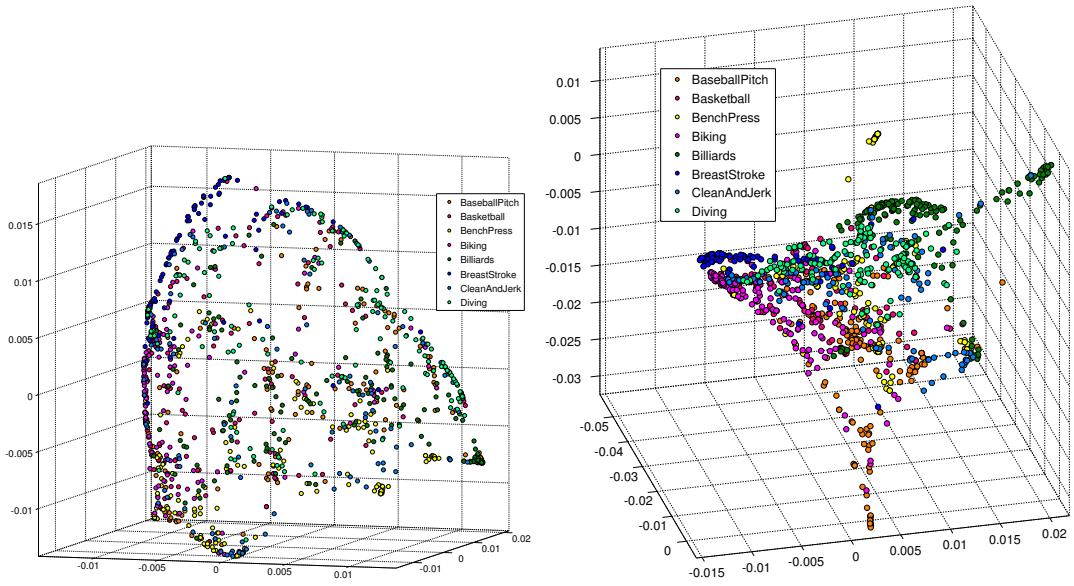


Figure 3.3: **Vector Space Mapping of Covariance Matrices from Appearance Cues**. The above two figures show covariance matrices (each matrix is a point) from video samples and their respective mapping in log-matrix space. In both the cases, representative video samples from 8 arbitrary classes in UCF50 are chosen and their respective covariance matrices are determined. Different classes are colored in differently. (a) shows original covariance matrices based on appearance features before mapping, and (b) shows the same after mapping.Note how some classes show more separability than others after the mapping.

As motion in a video can be characterized using simple temporal gradient (frame difference), horizontal (u) and vertical (v) components of optical flow vector, we use the following as our basic motion features:

$$\mathbf{f_m} = \begin{bmatrix} \frac{\partial I}{\partial t} & u & v & \frac{\partial u}{\partial t} & \frac{\partial v}{\partial t} \end{bmatrix}^T, \tag{3.2}$$

where $\frac{\partial}{\partial t}$ represents the finite differential operator along the temporal axis. In addition to these basic flow features, we extract high-level motion features [7] derived from concepts of fluid dynamics, since these are observed to provide a holistic notion of pixel-level motion within a certain spatial neighborhood. For e.g. features such as divergence ∇ and vorticity Γ quantify

the amount of local expansion occuring within flow elements and the tendency of flow elements to "spin", respectively. Thus

$$\begin{aligned} \nabla &= \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}, \\ \Gamma &= \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}. \end{aligned} \tag{3.3}$$

Local geometric structures present in flow fields can be well captured by tensors of optical flow gradients [7], which is mathematically defined as:

$$G = \begin{pmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{pmatrix}. \tag{3.4}$$

With this intuition, we compute the principal invariants of the gradient tensor of optical flow. These invariants are scalar quantities and they remain unchanged under any transformation of the original co-ordinate system. We determine the second, $\tau_2(G)$ and third $\tau_3(G)$ invariants of G as:

$$\begin{aligned} \tau_2(G) &= \frac{1}{2}\left[tr(G)^2 + tr(G^2)\right], \\ \tau_3(G) &= -det(G). \end{aligned} \tag{3.5}$$

Based on the flow gradient tensor, we determine the rate of strain, S and rate of rotation, R tensors which signify deviations from the rigid body motion, frequently seen in articulated human body movements. These are scalar quantities computed as :

$$
\begin{aligned}
S &= \frac{1}{2}(G + G^T), \\
R &= \frac{1}{2}(G - G^T).
\end{aligned} \tag{3.6}
$$

Using the equations in(3.5), principle invariants can be computed for these tensors. The interested reader is requested to read [7] for further insights on the selection of invariants. However, unlike the authors of [7], we do not compute the symmetric and asymmetric kinematic features as these assume human motion is centralized which is not valid for events occuring in an unconstrained manner (typically seen in YouTube videos). For the sake of legibility, we arrange the kinematic features computed from optical flow vectors in the following way,

$$
\mathbf{f}_k = \begin{bmatrix} \nabla & \Gamma & \tau_2(G) & \tau_3(G) & \tau_2(S) & \tau_3(S) & \tau_3(R) \end{bmatrix}^T. \tag{3.7}
$$

Finally we obtain the following representation for each pixel after concatenating all the above features to form a 19 element vector as:

$$
\mathbf{F} = \begin{bmatrix} \mathbf{f_i} & \mathbf{f_g} & \mathbf{f_m} & \mathbf{f}_k \end{bmatrix}^T. \tag{3.8}
$$

Fig. 3.1(b) visualizes the different low-level features that contribute to motion description from a sample action frame from UCF50. Fig. 3.4 provides a 3-dimensional visual illustration of covariance descriptors constructed using only the aforementioned motion cues on 8 different action classes in UCF50 and their equivalent vector space representations.

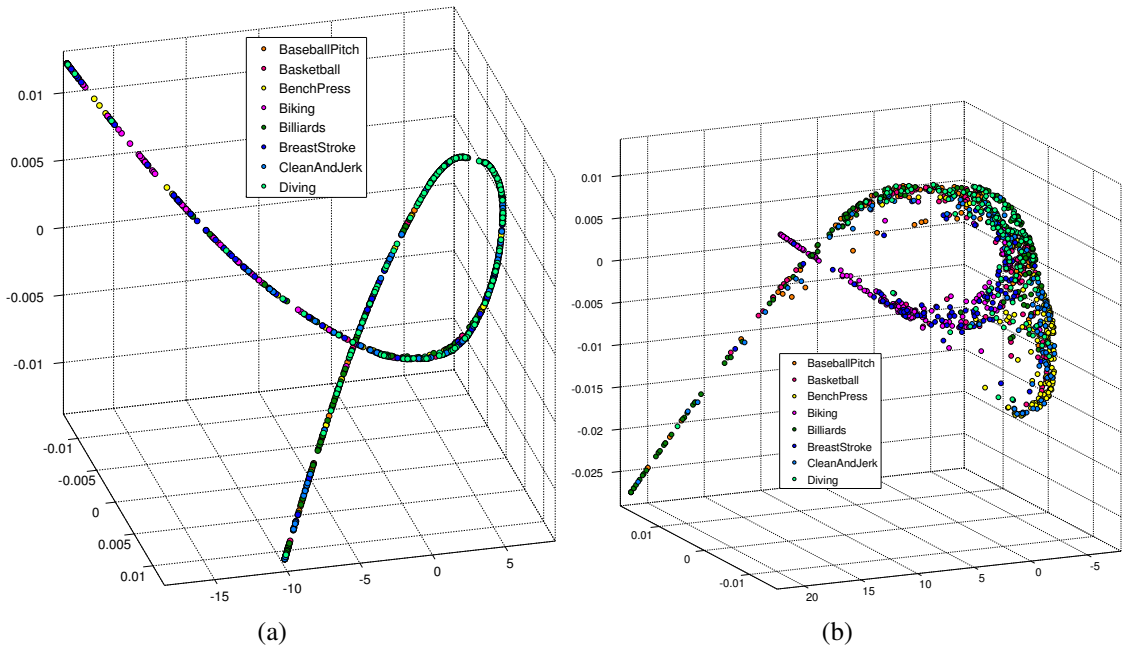


Figure 3.4: **Vector Space Mapping of Covariance Matrices from Motion Cues**. The above two figures show covariance matrices (each matrix is a point) from video samples and their respective mapping in log-matrix space. In both the cases, representative video samples from 8 arbitrary classes in UCF50 are chosen and their respective covariance matrices are determined. Different classes are colored in differently. (a) shows original covariance matrices based on motion features before mapping, and (b) shows the same after mapping.Note how some classes show more separability than others after the mapping.

3.3 Feature Fusion using Covariance Matrix

Covariance based features introduced by Tuzel and colleagues for object recognition [163] have found application in various other related areas such as: face recognition [125, 146], shape modeling [174], and object tracking [134]. Based on an integral image formulation as proposed in [163], we can efficiently compute the covariance matrix for a video clip where each pixel is a sample. The covariance matrix in this context is therefore computed as :

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{f_i^{(k)}} - \mu)(\mathbf{f_i^{(k)}} - \mu)^T, \tag{3.9}$$

where $\mathbf{f^{(k)}}$ is a single feature set and μ is its corresponding mean, n being the number of

samples (here pixels). Since the covariance matrix is symmetric, it contains $(d^2 + d)/2$ (d being the total types of features) unique entries forming the upper or lower triangular part of the matrix, that capture cross feature set variance.

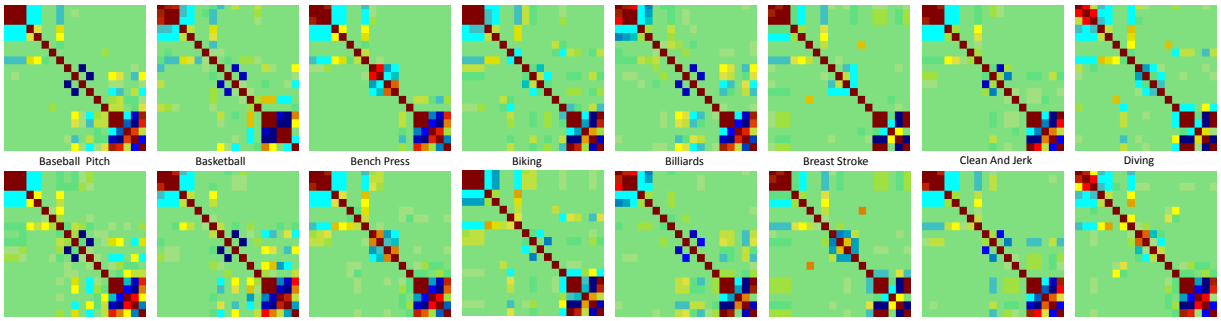


Figure 3.5: **Normalized Covariance matrices from** 8 **class of actions from UCF50:** Each column shows a different class, and each row is a sample covariance matrix constructed from clips belonging to one of the 8 classes. We can notice the subtle differences between two samples of different classes and some structural similarity of elements of the same class. This aspect is more pronounced in Fig. 3.6.

Covariance matrices have some interesting properties which naturally suits our problem. Since these matrices do not have any notion of the order in which samples are collected, they are computationally more favorable compared to trajectory based descriptors [170] that require explicit feature tracking. Secondly, covariance based descriptors provide a better way of analyzing relationship across feature sets compared to mere concatenation of histograms of different features [81]. Furthermore, the covariance matrices provide more concise representation of the underlying feature distribution due to symmetry compared to long descriptors generated by methods proposed in [39, 145] which need additional dimensionality reduction.

That said, addition and scalar multiplication on covariance matrices are not defined as these matrices conform to non-linear connected Riemannian manifolds of positive definite matrices (S_n^+). Hence, these matrices cannot be used as they are for classification using regular machine learning approaches that make assumptions on the data belonging to a linear subspace (unless special considerations are undertaken to understand the underlying feature space). One possible approach to address this issue is to map these matrices to an equivalent vector space closed under

addition or scalar multiplication, in order to facilitate classification tasks. However, in doing so, the structure of the covariance matrix which conforms to Riemannian geometry, is not exploited in the classification purpose. In view of this, we employ a sparse representation technique suited for covariance matrices, called tensor sparse coding [146]. In the next few sections we provide a brief theoretical background on our sparse representation framework for classification based on covariance matrices with empirical evaluations of the two methods we presented in this chapter.

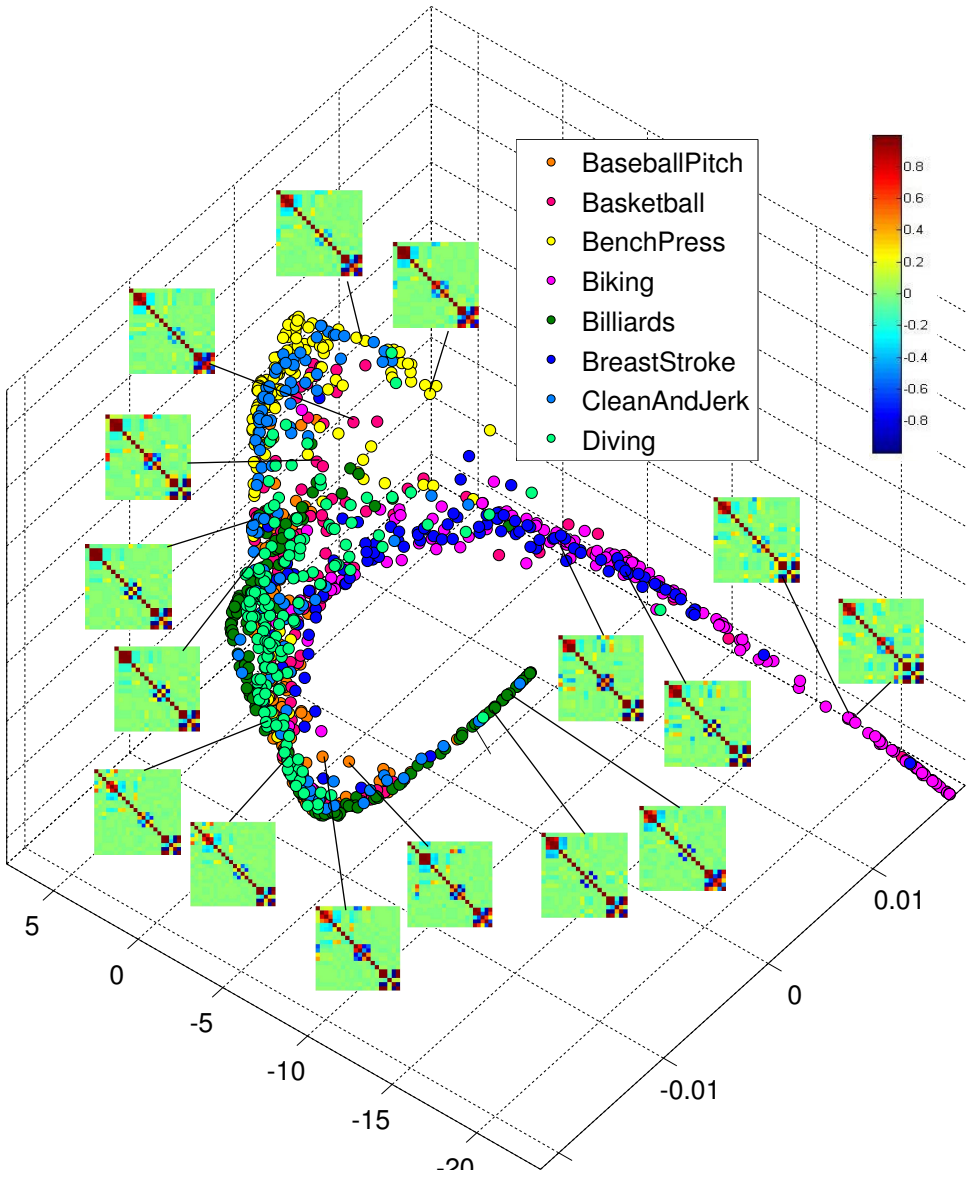


Figure 3.6: **Vector space mapping of covariance matrices** based on appearance and motion features. See Figs. 3.3-3.4 for detailed interpretation. Sample covariance matrices (as in Fig.3.5) are shown as insets.

3.4 Classification using Sparse Representation

Recently, sparse linear representation techniques have shown promising results in solving key computer vision problems including face recognition [183], object classification [105] and action recognition [55]. The basic objective of these approaches is to project the classification problem into a sparse linear approximation problem. Formally, given a set of K training samples consisting of k classes, $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_K$ and a test sample $\mathbf{y}$, an overcomplete dictionary A is constructed by stacking the training samples. Then the approximation problem:

$$\min ||\mathbf{x}||_1 \quad s.t. \quad \mathbf{y} = A\mathbf{x} \tag{3.10}$$

where $\mathbf{x}$ is a sparse vector of coefficients corresponding to each element in A, can be solved using linear programming techniques. For each coefficient in $\mathbf{x}$, the residuals :

$$r_i = ||\mathbf{y} - A\mathbf{x}_i||_2 \tag{3.11}$$

are computed, where $\mathbf{x_i}$ is a zero vector with ith entry set to the ith coefficient in $\mathbf{x}$. The smallest residual identifies the true label of $\mathbf{y}$.

This classification strategy, although computationally attractive, cannot be directly applied to our problem as it expects the samples to span vector spaces (linear manifolds in R^2). Fortunately, such an equivalent vector space for positive definite matrices exists, where these matrices can be mapped to the tangent space of the Riemannian manifolds [10]. There are a couple of advantages of using this transformation, besides of the utility of being used in linear classification algorithms. The distance metric defined in this transformed subspace, is affine invariant and satisfies triangle inequality. The interested reader is requested to refer to [50] to get theoretical proofs. These properties make our descriptor robust to a certain degree of noise in the data. Having said that, we

perform this transformation of a covariance matrix C to its log L using :

$$L = \log(C) = R^T \hat{D} R, \tag{3.12}$$

where R^T, R are rotation matrices obtained after singular value decomposition of C and $\hat{D}$ is the diagonal matrix containing the log of eigenvalues obtained after SVD. The mapping is a symmetric matrix whose upper or lower triangular components form our final feature descriptor for a given video clip.

3.4.1 Sparse Coding of Matrix Log Descriptors

Given a set of clips from training videos, we construct our overcomplete dictionary (A) consisting of their p corresponding matrix log descriptors. For a query video of m clips, there are m matrix log descriptors. Since $p >> m$, the original sparse linear approximation needs to be modified, to obtain a computationally tractable solution. We use an efficient implementation [1] of the Orthogonal Matching Pursuit [159] algorithm to achieve this. The algorithm solves the sparse approximation problem where coefficients are constrained to be the orthogonal projection of the query sample $\mathbf{y}_j$ on the dictionary A. In this context, the problem can be stated as :

$$\min || \{\mathbf{y}_j\}_{j=1}^{m} - A\mathbf{x}||_2^2 \quad s.t. \quad ||\mathbf{x}||_0 \leq P, \tag{3.13}$$

with $||\mathbf{x}||_0$ the L_0 pseudo-norm equal to the number of nonzero coefficients in $\mathbf{x}$, P being an empirically determined threshold ensuring the degree of sparsity. The solution to this problem provides a set of m labels corresponding to each clip from the query video. The final label of the video can be obtained using a simple majority voting.

The technique discussed above can be viewed as a simple straight-forward solution to our problem as there have been limited research to extend this idea towards matrices or tensors. This

[1] http://www.cs.technion.ac.il/∼ronrubin/Software/ompbox10.zip

motivates us to explore further on the recent advances of Sivalingam and colleagues [146] in sparse coding of covariance matrices which is discussed as follows.

3.4.2 Tensor Sparse Coding of Covariance Matrices

In order to address the problem of sparse linear approximation of covariance matrices, we begin with the following formulation: Consider our query video consists of a single clip whose motion-appearance covariance matrix Q, constructed using Eqn. 3.9, can be expressed as a linear combination of covariance matrices forming an overcomplete dictionary D:

$$Q = x_1 D_1 + x_2 D_2 + \ldots + x_p D_p = \sum_{i=1}^{p} x_i D_i, \tag{3.14}$$

where x_i's are coefficients of the elements D_i from dictionary D of covariance matrices of labeled training videos. As Q belongs to the connected Riemannian manifold of symmetric positive definite matrices, the following constraint is implied:

$$\hat{Q} \succeq 0, \Rightarrow x_1 D_1 + x_2 D_2 + \ldots + x_p D_p \succeq 0, \tag{3.15}$$

where $\hat{Q}$ is the closest approximation of Q, introduced to handle noise in real-world data. This closest approximation can be achieved by solving an optimization problem. However, in order to perform this task, we first need to define a measure of proximity between our query matrix Q and the approximated solution $\hat{Q}$. Such a proximity measure is often measured in terms of penalty function called LogDET or Burg matrix Divergence [64] which is defined as:

$$\Phi_{\nabla}(\hat{Q}, Q) = tr(\hat{Q}Q^{-1}) - \log\det(\hat{Q}Q^{-1}) - d, \tag{3.16}$$

Using Eqn.(3.14), the above equation can be further expanded as:

$$\Phi_{\nabla}(Q, \hat{Q}) = tr(\sum_{i=1}^{p} x_i D_i Q^{-1}) - \log\det(\sum_{i=1}^{p} x_i D_i Q^{-1}) - d, \tag{3.17}$$

Since, $\hat{D}_i = Q^{-1/2} D_i Q^{-1/2}$, we can substitute Eqn.(3.17) appropriately, achieving:

$$\begin{aligned} \Phi_{\nabla}(Q, \hat{Q}) &= tr(\sum_{i=1}^{p} x_i \hat{D}_i) - \log\det(\sum_{i=1}^{p} x_i \hat{D}_i) - d, \\ &= \sum_{i=1}^{p} x_i tr(\hat{D}_i) - \log\det(\sum_{i=1}^{p} x_i \hat{D}_i) - d, \end{aligned} \tag{3.18}$$

where the $\log\det(.)$ function can be expressed as Burg Entropy of eigenvalues of a matrix Z as $\log\det(Z) = \sum_i \log\lambda_i$. Therefore, our optimization problem can be formulated using the objective function in Eqn.(3.18) as:

$$\begin{aligned} \min_x \quad & \sum_{i=1}^{p} x_i tr(\hat{D}_i) - \log\det(\sum_{i=1}^{p} x_i \hat{D}_i) + \delta||\mathbf{x}||_1 \\ \text{subject to} \quad & \mathbf{x} \geq 0, \quad \sum_{i=1}^{p} x_i \hat{D}_i \succeq 0, \text{and}, \quad \sum_{i=1}^{p} x_i \hat{D}_i \preceq I_n \end{aligned} \tag{3.19}$$

with, $\delta||\mathbf{x}||_1$ being a relaxation term that incorporates sparsity. The above problem can be mapped to a determinant maximization problem which can be efficiently solved by semi-definite programming techniques. We use the sparse approximation toolbox[2]. Multiple covariance matrices per video are handled in a similar way as mentioned in the case of sparse coding with matrix logs. In the next sections, we provide our experimental details comparing the approaches presented here on two different application domains.

[2]http://cvxr.com/cvx/

3.5 Action Recognition using Covariance Descriptors

This is an extremely challenging problem, especially because videos depicting events are captured in diverse settings. There are two newly introduced, challenging datasets (UCF50 [3], HMDB51 [80]) containing videos that reflect such settings (multiple and natural subjects, background clutter, jittery camera motion, varying luminance). To systematically study the behavior of our proposed descriptor and the associated classification methods, we conduct preliminary experiments on a relatively simple, well recognized, human actions dataset [144] to validate our hypothesis and then proceed towards the unconstrained case.

3.5.1 Datasets

KTH Human Actions: This dataset [144] consists of 6 classes namely: Boxing, Clapping, Jogging, Running, Walking, and Waving. The dataset is carefully constructed in a restricted environment – clutter-free background, exaggerated articulation of body parts not seen in real-life, mostly stable camera except for controlled zooming with single human actors. The videos in this dataset are in gray scale and not much cue is useful from background.

UCF50: The UCF50, low-level event dataset [3] consists of video clips that are sourced from YouTube videos (unedited) respectively. It consists of over $6,500$ RGB video clips (unlike KTH) distributed over 50 complex human actions such as horse-riding, trampoline jumping. baseball pitching, rowing etc. This dataset has some salient characteristics which makes recognition extremely challenging as they depict random camera motion, poor lighting conditions, huge foreground and background clutter, in addition to frequent variations in scale, appearance, and view points. To add to the above challenges, since most videos are shot by amateurs with poor cinematographic knowledge, often it is observed that the focus of attention deviates from the foreground.

HMDB51: The Human Motion DataBase [80], introduced in 2011, has approximately $7,000$ clips distributed over 51 human motion classes such as : brush hair, push ups, somersault etc. The videos have approximately 640×480 spatial resolution, and are mostly sourced from TV shows and movies. The videos in the dataset are characterized by significant background clutter, camera jitter and to some extent the other challenges observed in the UCF50 dataset.

3.5.2 Experimental Setup

We make some adjustments to the original covariance descriptor by eliminating appearance based features in Eqn.(3.8) to perform evaluations on the KTH dataset, as not much contextual information is available in this case. Thus each pixel is represented by a 12 dimensional feature vector (last 12 features from $\mathbf{F}$ in 3.8) resulting in a $(12^2 + 12)/2 = 78$ dimensional vector. Each video is divided into uniformly sampled non-overlapping clips of size $w \times h \times t$, w, h being the original resolution of the video and t is the temporal window. Throughout all experiments, we maintain $t = 20$. Optical flow which forms the basis of our motion features, is computed using an efficient GPU implementation [31].

For all classification experiments we use a split-type cross-validation strategy suggested by the authors in [144]. We ensure that the actors that appear in the validation set do not appear in the training set to construct a dictionary for fair evaluation. Similar split strategy is employed for experiments on UCF50. For HMDB51 we follow the authors validation strategy that has three independent splits. The average performance across all splits is recorded in Tables 3.1 and 3.2.

We compare the proposed sparse representation framework against a linear SVM [46] classifier that uses the matrix-log descriptors from a video as feature vectors. Descriptors from each clip is treated independently to obtain initial class labels and latter using a simple majority voting mechanism, the respective votes are fused to determine the true class label. This voting strategy is

kept same for both the proposed sparse representation based classification schemes as well.

3.5.3 Results

In this section, we take the opportunity to summarize our observations under different experimental setups. To investigate the contribution of different feature modalities towards the recognition performance, we computed 3 different sets of covariance matrices for videos in UCF50. Firstly, descriptors computed using only appearance features (resulting in a 7×7 matrix). Next, we use only motion based features. Thus the covariance matrix in this case is 12×12. Finally, both appearance and motion features are used together to compute the covariance matrices. We also evaluated how each classification strategy behaves with these different descriptors. For each of these descriptors, the classification framework was varied between a linear SVM (LC/SVM), Sparse OMP (LC/OMP), and finally the Tensor Sparse Coding (TSC) algorithm that uses MAXDET optimization. For the first two methods, the descriptors are thus 28 (appearance), 78 (motion) and 190 (all).

Individual Feature Contributions: We observed that the appearance features are less informative as compared to the motion features in videos where RGB information is available. However, all classification techniques get a boost in performance when both the features are used together.

Tensor Sparse Coding based classification performs better than other two methods. Among classifiers, linear SVM and OMP, we observe OMP perform better than the former which shows that there is an inherent sparsity in the data which is favored by sparse representation based classification techniques. Table 3.1 summarizes the results of the experiments involving the contribution of different feature modalities and methods. The different columns in the table show the feature modalities used for computing the covariance matrices (AF = Appearance Features, MF = Motion Features, AMF = Appearance and Motion Features).

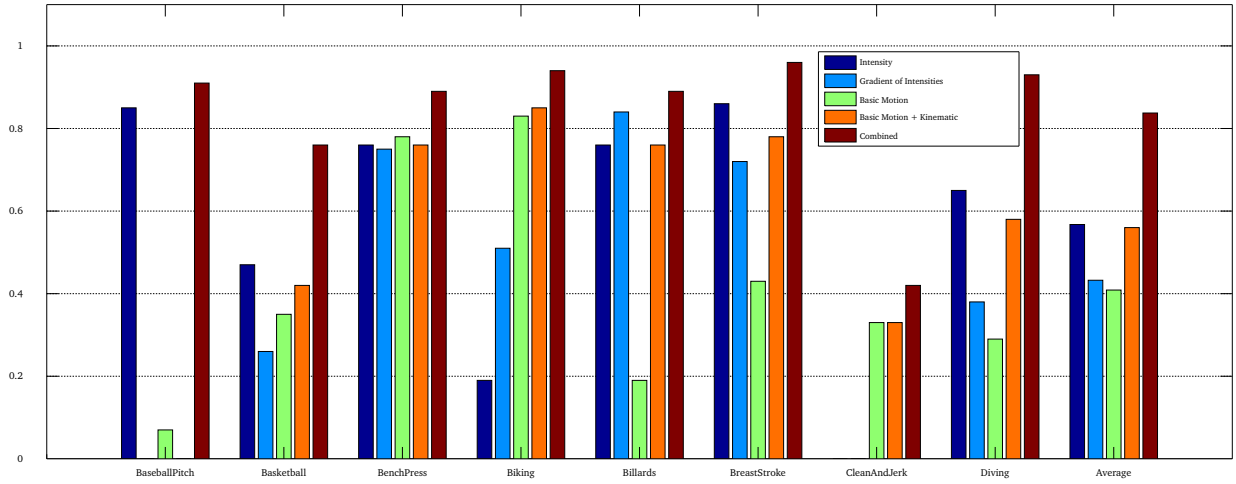
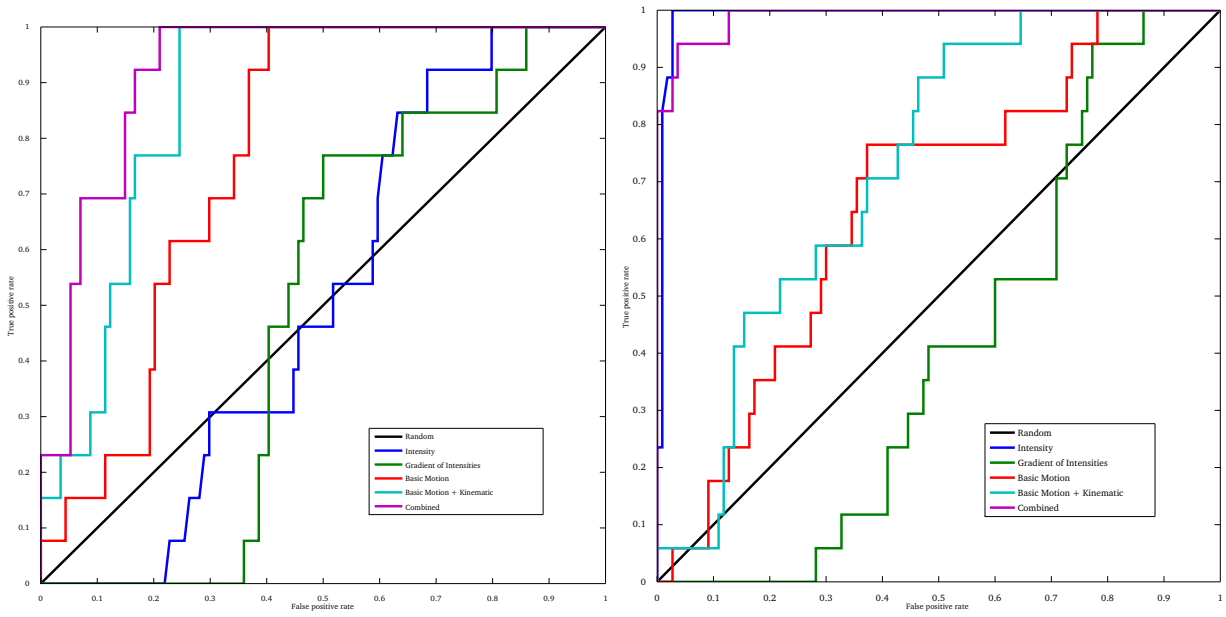


Figure 3.7: **F-measures for** 8 **classes from UCF50 dataset with different features:** The features experimented with are as follows : $\mathbf{F_i}$ (RGB Intensities), $\mathbf{F_g}$ (Intensity gradients), $\mathbf{F_m}$ (Basic Motion Features i.e. temporal derivatives, optical flow etc.), $\mathbf{F_k}$ (Kinematic Features), A combination of basic motion and kinematic features and finally all features are used together in the covariance descriptor.

In order to provide a more detailed insight on the individual feature contribution towards the overall classification performance, we experiment with different features in a finer granularity. The Fig.3.7 indicates F-measures derived from precision and recall for 8 different classes of unconstrained actions from UCF50 dataset. It is interesting to notice two distinct trends from this experiment: RGB intensities contribute the most towards the discriminativity of the covariance descriptor for "Baseball-pitch class while the "CleanAndJerk" is best described by motion features. This can be explained by the sudden vertical motion captured by the basic motion and kinematic features in "CleanAndJerk" samples, and the mostly greener texture of background captured by intensity features in "Baseball-pitch samples. The ROC curves for detection of these classes are provided in Fig. 3.8, emphasize the contribution of the features in further finer granularity.



(a) (b)

Figure 3.8: **ROC curves for detection** of (a) "CleanAndJerk" and (b) "Baseball-pitch" samples from UCF50. Of the 8 classes analyzed in Fig. 3.7, these are the two classes which have clear separation because of their distinctive motion features.

In Table 3.2, we present a comparative analysis of the various classification methods on these datasets. We compare our methods [20] with the state of the art performance obtained by other competitive approaches. Although our proposed method does not show any improvement over the state of the art on the KTH dataset, we observe definite increase in performance over the two other complex event recognition datasets. We also observe that there is a steady increase in performance across the datasets as we change our classification strategies from a linear SVM to the more complex tensor sparse coding (TSC) based classification scheme. Note that the performance reflected in case of UCF50 and HMDB51 datasets are significantly high as compared to other approaches. However, it is to be noted that the MAXDET optimization is a computationally intensive operation which forms the basis of TSC.

Table 3.1: **Contribution of feature sets and methods:** This table summarizes the impact of different feature sets and different methods on our experiments in UCF50. The rows indicate different methods: LC/SVM (Matrix Log descriptors from covariances with linear SVM classifier), LC/OMP (Same descriptor, sparse OMP classifier), and TSC (Tensor Sparse Coding of Covariance Matrices) columns show the feature modalities used for computing the covariance matrices (AF = Appearance Features, MF = Motion Features, AMF = Appearance and Motion Features).

	Experiments on UCF50		
Method	**AF**	**MF**	**AMF**
LC/SVM	31.4%	43.4%	47.4%
LC/OMP	34.2%	42.5%	51.5%
TSC	34.5%	46.8%	53.8%

Table 3.2: **Comparison with the state-of-the-art methods:** This table summarizes the performance of two of our proposed methods [20] with respect to already published state-of-the-art approaches. The first row is used as a baseline (BL) where we use our implementation of a bag-of-visual-words based representation on a good local feature descriptor (HOG-HOF proposed by Laptev and colleagues [81]. The second row i.e, LC/SVM uses a linear SVM as a classifier on top the matrix log descriptors while the bottom two rows use the OMP and MAXDET based sparse coding for classification.

	Datasets		
Method	**KTH**	**UCF50**	**HMDB51**
BL [81]	92.0%	48%	20.2%
LC/SVM	86.2%	47.4%	21.03%
LC/OMP	88.2%	51.5%	22.09%
TSC	91.4%	53.8%	26.16%

Finally, in Fig. 3.9 and Fig. 3.10, we present the confusion matrices obtained after classification using the tensor sparse coding which performs the best in case of both the datasets. In UCF50, the highest accuracies are obtained for classes that have discriminative motion (e.g. Trampoline jumping is characterized by vertical motion as opposed to other categories). The following section provides a brief discussion on the algorithmic complexities involved in the various steps of the entire recognition pipeline.

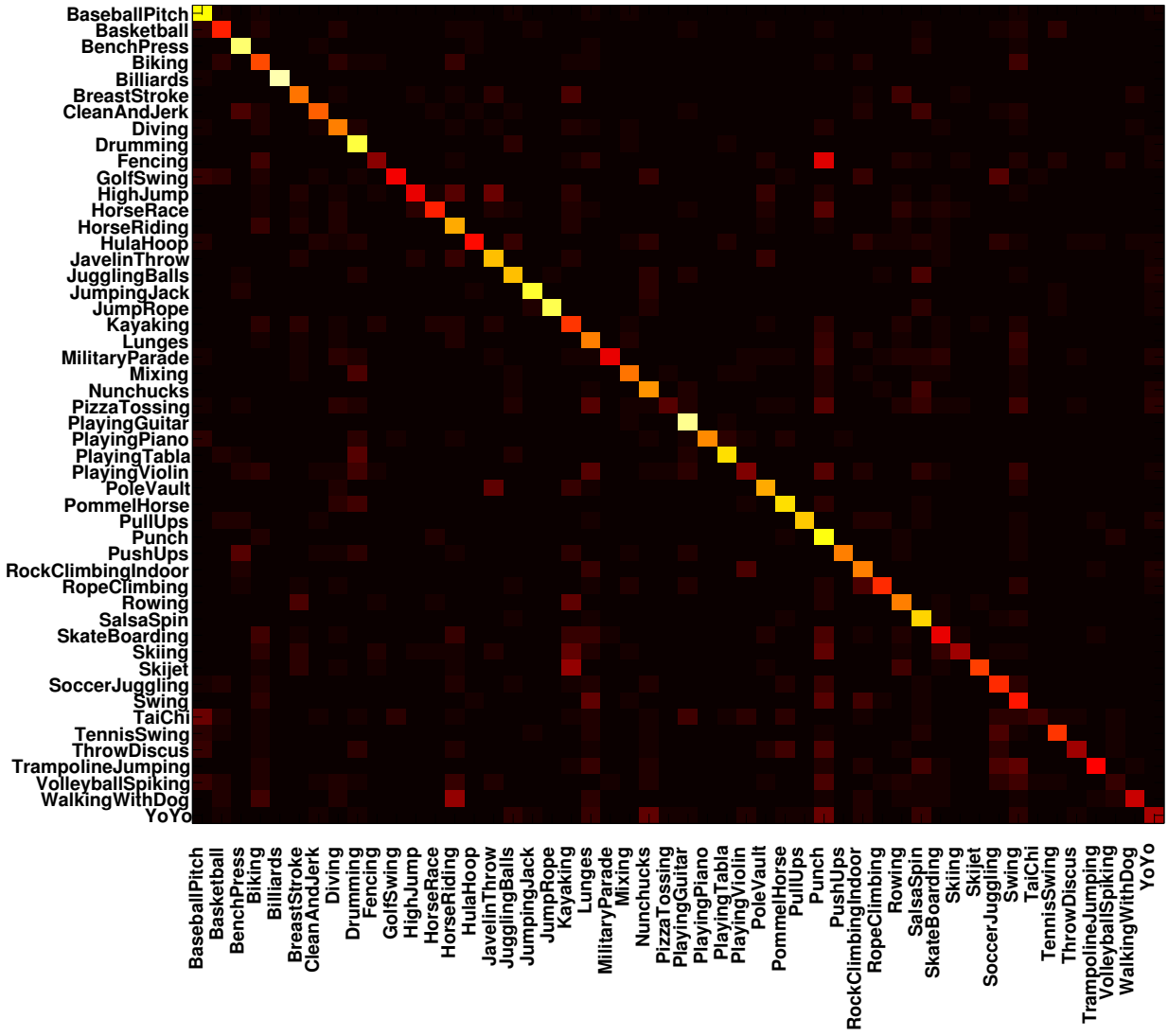


Figure 3.9: **Confusion matrix** obtained after performing classification using the proposed classification technique on the UCF50.

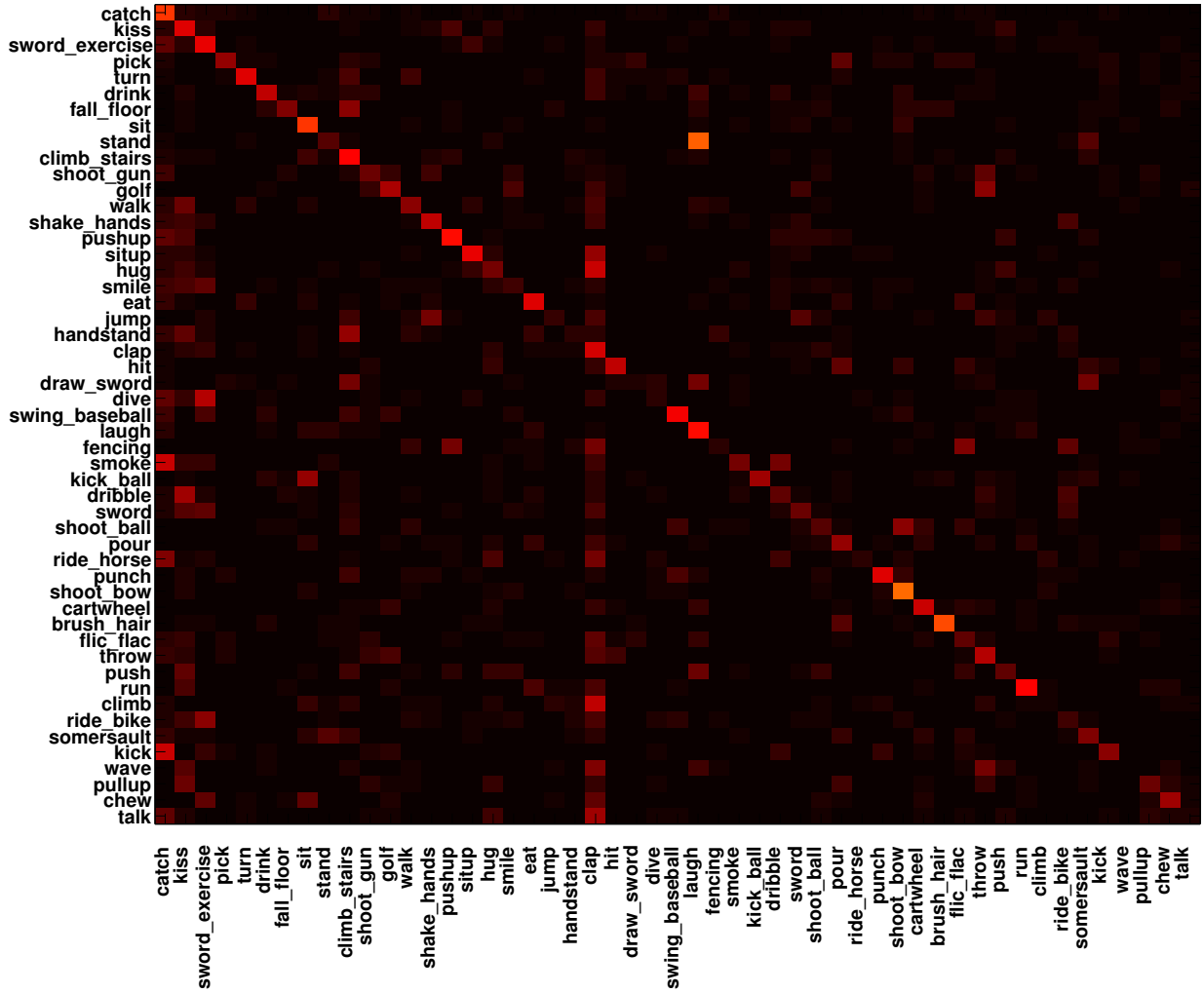


Figure 3.10: **Confusion matrix** obtained after performing classification using the proposed classification technique on HMDB51 dataset.

3.5.4 Complexity Analysis

The entire computation pipeline can be summarized in three major steps, namely low-level feature extraction, feature fusion using covariance matrices, followed by classification. Off these, the feature extraction and covariance computation step for each clip of a video can be done in parallel for any dataset. Among feature extraction, optical flow computation [31] is the most expensive step, which is based on a variational model. For a consecutive pair of frames, with a

resolution of 512×384, a GPU implementation of the above algorithm, takes approximately 5 seconds on a standard desktop computer hosting a 2.2Ghz CPU with 4GB of physical memory. Depending on the types of low-level features computed, the complexity of the covariance matrix computation is $O(WHdC)$ where d is the total types of low-level features, W and H are the respective width and height of a typical clip, and C being the total number of frames per clip.

The complexity of classification using the Orthogonal Matching Pursuit [159] scheme is optimized using an efficient batch implementation provided in [140]. Since this method involves precomputation of an in-memory dictionary of fixed number of elements (T_D), the overall complexity can be approximated as $O(T_D + K^2d + 3Kd + K^3)$, where K is the target sparsity for sparse coding. For details, please refer [140]. Classification using MAXDET optimization, on the other hand, is relatively more expensive as it attempts to find a subset of dictionary atoms representing a query sample using a convex optimization. In closed form, this is $O(d^2L^2)$, L being the number of dictionary atoms. Although, this technique is more reliable in terms of accuracy, it requires a larger computation overhead as the process needs to be repeated for every query sample. Assuming the number of samples are far larger than L batch-OMP is observed to offer a respectable trade-off between accuracy and speed.

3.6 One-shot Learning of Human Gestures

In addition, to demonstrate the applicability of our video descriptor, we report our preliminary experimental results on different application domain: human gesture recognition using a single training example.

3.6.1 ChaLearn Gesture Data (CGD) 2011

This dataset is compiled from human gestures sampled from different lexicons e.g. body language gestures (scratching head, crossing arms etc.), gesticulations performed to accompany

speech, sign languages for deaf, signals (referee signals, diving signals, or marshaling signals to guide machinery or vehicle) and so on. Within each lexicon category, there are approximately, 50 video samples organized in different batches, captured using depth and RGB sensors provided by the Kinect [3] platform. Each video is recorded at 30 Hz at a spatial resolution of 640×480. Each batch is further divided into training and testing splits and only a single example is provided per gesture class in the training set. The objective is to predict the labels for the testing splits for a given batch.

Although the videos are recorded using a fixed camera under homogeneous lighting and background conditions, with a single person performing all gestures within a batch, there are some interesting challenges in this dataset. These are listed as follows: (1) Only one labeled example of each unique gestures, (2) Some gestures include subtle movement of body parts (numeric gestures), (3) Some part of the body may be occluded, and, (4) Same class of gesture can have varying temporal length across training and testing splits.

[3] http://en.wikipedia.org/wiki/Kinect

Figure 3.11: Sample frames from representative batches from the CGD 2011 dataset.

3.6.2 Experimental Setup

We obtain a subset of 10 batches from the entire development set to perform our experiments. For a given batch, the position of the person performing the gesture remains constant, so we adjust our feature vector in Eqn.(3.8) to incorporate the positional information of the pixels x, y, t in the final descriptor. Furthermore, since the intensities of the pixels remain constant throughout a given batch, the RGB values at the corresponding pixel locations could also be eliminated. Also, the higher order kinematic features such as $\tau_2(S)$, $\tau_3(S)$, and $\tau_3(R)$ can be removed as they do not provide any meaningful information in this context. Thus each pixel is represented in terms of a 16 dimesional feature vector, resulting in a 16×16 covariance matrix with only 136 unique entries. The upper triangular part of the log of this matrix forms our feature descriptor for a clip extracted from a video. In order to perform classification, we use a nearest neighbor based classifier with the same clip-level voting strategy as discussed in the earlier experiments. A regular SVM based classifier is not applicable to this problem as there is only one training example from each gesture class.

Table 3.3: Comparison with other methods: This table summarizes the performance of our descriptor in one-shot gesture recognition against other methods. The leftmost column contains the batches on which the methods are tested. The next two columns contain indicate the avg. accuracy obtained using two local feature based approaches: MBH [170] and STIP [81] using a Nearest Neighbor (NN) classifier, the next column uses template matching based method and the last column records the performance of our descriptor (LC) when used with a nearest neighbor classifier.

	Method Acc. Avg. (%)			
Batch ID	MBH/NN	STIP/NN	TPM	LC/NN
Devel01	66.7	25.0	58.3	83.3
Devel02	33.4	8.3	25.0	75.0
Devel03	7.2	28.6	14.3	28.6
Devel04	33.4	16.7	58.4	75.0
Devel05	28.6	14.3	64.3	100.0
Devel06	50.0	16.7	25.0	91.7
Devel07	23.1	7.7	15.4	84.6
Devel08	36.4	9.09	9.1	81.8
Devel09	30.7	23.1	53.8	69.2
Devel10	15.4	15.4	23.1	53.6
Avg.	32.4	16.4	34.7	74.3

Since depth information is available along with the RGB videos, we exploit it to remove noisy optical flow patterns generated by pixels in the background, mainly due to shadows.

3.6.3 Results

Similar to the previous experiments on low-level event recognition in section 3.5.3, we perform a detailed analysis, with more emphasis on the descriptor. To this end, we use different versions of the descriptor with only motion features (M:9×9 covariance matrix), a combination of motion and intensity gradients (MG:13×13 covariance matrix), a combination of motion and positional information (MP:12×12 covariance matrix) and finally all features combined (16×16). The results are reported in Table 3.4. We observe that again motion in itself is not the strongest cue. However, when fused with appearance gradients and positional information, the overall performance of the descriptor increases by 11%, which is a significant improvement considering the nature of the problem.

Table 3.4: **Contribution of different low level features towards the one-shot gesture learning problem:** Each column shows a different set of low-level features used to compute the final descriptor. The order of low-level feature sets are as follows: Basic Motion (M), Basic Motion and Intensity gradients (MG), Subset of Basic motion and positional informations (MP), and finally all features combined. Refer to Section 3.6.2 for more details .

	Descriptor Performance(%)			
Batch ID	M	MG	MP	All
Devel01	66.7	66.7	88.3	83.3
Devel02	53.3	66.7	53.3	75.0
Devel03	28.6	42.9	21.4	28.6
Devel04	53.3	58.3	75.0	75.0
Devel05	92.8	100	92.8	100.0
Devel06	83.3	91.7	83.3	91.7
Devel07	61.5	76.9	61.5	84.6
Devel08	72.7	72.7	81.8	81.8
Devel09	69.2	61.5	69.2	69.2
Devel10	38.5	61.5	53.6	53.6
Avg.	62.9	69.9	68.0	74.3

In order to make a fair evaluation of our descriptor with the state-of-the-art descriptors from action recognition literature [81, 170], we keep the classifier constant (Nearest Neighbor). We also compared our approach with a simple template matching based recognition which is more appropriate for this type of problem. The average accuracies for each batch tested using all the compared methods are reported in Table 3.3. It is pleasing to note that our descriptor performs significantly better than all other methods which gives us promising leads towards the applicability of this descriptor for this class of problems. Finally, in Fig. 3.12, we show the respective confusion matrices obtained after applying the proposed method on first 10 of the development batches from the CGD 2011 dataset.

3.7 Summary

In this chapter, we presented an end-to-end framework [20] for action and gesture recognition. As part of this effort, we introduced a novel descriptor for general purpose video analysis

that can be considered as an intermediate representation between local interest point based feature descriptors and global descriptors. We described how simple appearance and motion cues can be efficiently integrated to form covariance descriptors, that efficiently encodes meaningful second order statistics of the data. We also proposed two sparse representation based classification approaches that can be applied to our descriptor.

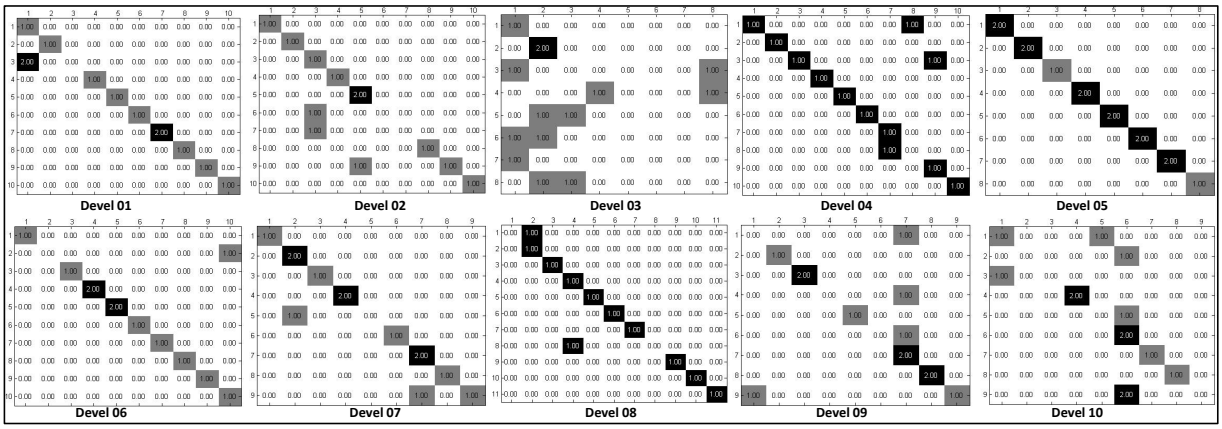


Figure 3.12: Confusion matrices obtained after applying the proposed method on first 10 of the development batches from the CGD 2011 dataset. Note for certain batches (devel01–02, devel04–05, devel06–08), our method is able to predict gesture labels with respectable accuracies using just one training sample.

CHAPTER 4: CINEMATOGRAPHIC SHOT CLASSIFICATION AND ITS APPLICATION TO COMPLEX EVENT RECOGNITION

4.1 Introduction

Research in computer vision and multimedia is constantly exploring novel information modalities to facilitate extraction of meaningful yet discriminative features. One such possible exploit is the inherent camera motion that is often prevalent during capture process. In this chapter, we introduce a novel stack of methodologies [19], that can be used to recover this motion and used in an effective way to perform some visual recognition tasks. Our technique derives its inspiration from cinematographic principles.

In cinematographic terminology [9], an authored video consists of several hierarchical components. An authored video can be divided into a collection of disjoint components called *scenes*. Each scene consists of a set of *shots* that are a collection of *frames*, which are at the lowest level of the hierarchy. Most video analysis algorithms take shots as input as they provide an intermediate yet rich representation of a video both in terms of its content and the motion of the camera during the shooting process, which usually adhere to certain production grammars. Shot level classification of videos has been an interesting field in computer vision research, especially due to its applicability to diverse domains. These include content based video search [148], film genre classification [137, 139] and video quality analysis. With Internet users becoming more and more selective about the results returned by today's video search websites, there is a pressing need to pursue classification of videos from different perspectives. There has been a cornucopia of research [40, 45, 111, 116, 117, 137, 139, 177] that address shot classification exploiting various low-level features such as textures, intensity etc. While these features are meaningful at content level, they are unable to capture the ambient camera motion which replicates the narrative human eye and hence are far more semantically challenging.

Camera motion in authored videos (commonly pan, tilt or zoom), are directly correlated with high-level semantic concepts described in the shot. For e.g, a *tracking shot* in which a camera undergoes translation on a moving platform indicates the presence of a *following* concept. Detection of such useful concepts can be used by current video search engines at a later stage to perform high-level content analysis such as detection of events from videos. This motivates us to explore the possibilities of using pure camera motion to solve the shot classification problem. Camera motion parameters, also known as telemetry, are very difficult to obtain directly as few video cameras are equipped with sophisticated sensors that can provide such accurate measurements. Furthermore, telemetry data is not generally available and is certainly not present in Internet or broadcast video. Hence, we resort to a purely image based technique to robustly estimate homographies which are coarse indicators of the camera motion incurred during capture. However, homographies are not meaningful features for discriminative classification of shots as different parameters in a homography matrix quantify different planar relationship (scale, rotation, etc.) and cannot be treated in separation. Also, since homographies belong to projective group (not closed under vector subtraction or scalar multiplication), they are not suitable for classifiers (linear kernel SVMs or Nearest Neighbors etc.) Therefore representing the ambient motion in a principled manner is extremely important, in order classify a shot.

While there exist methods [150, 181] to estimate camera motion using full 3D reconstruction of a scene, we argue that our method achieves a reasonable trade-off between high-accuracy and prohibitive computational cost. This enables us to contribute a global feature based on camera motion which can be used for large scale video analysis.

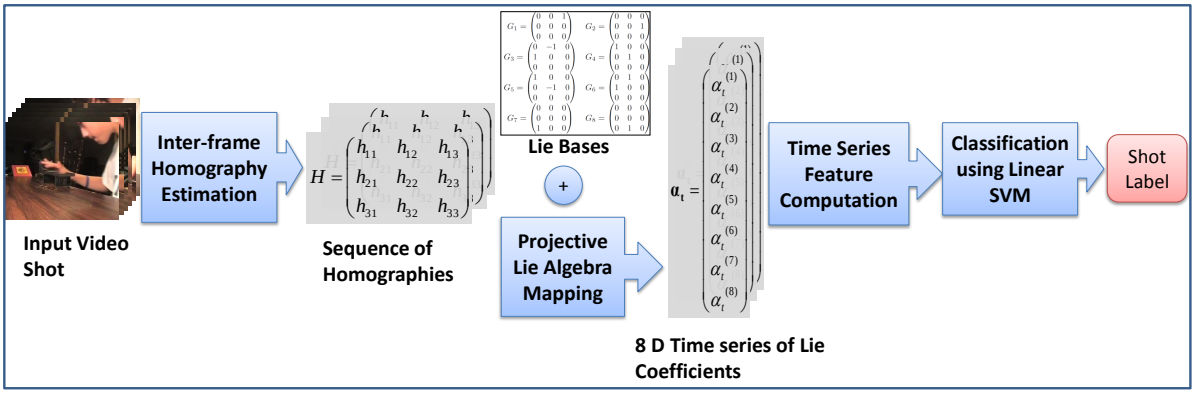


Figure 4.1: A schematic diagram showing the various processes involved in our proposed approach [19] towards classification of a typical shot. We build our complex event recognition computational pipeline (discussed in Section 4.7 based on the above methodology. Please refer to text for a detailed explanation.

To this end, we propose the following methodology (Fig. 4.1) to represent the camera motion extracted from a video: (1) Given a shot, pairwise homographies are computed between the consecutive frames, (2)Next we map them to a linear space using Lie algebra defined under Projective Group (3) Coefficients of this linear space are used to construct multiple time series (4) Representative features are computed from these time series for discriminative classification. A schematic diagram of our computational pipeline is shown in Fig. 4.1.

4.2 A Cinematography Primer

A complete list of cinematographic techniques can be found in [9]. In this chapter we focus on the following cinematographic shot classes:aerial, bird-eye, crane, dolly, establishing, pan, tilt and zoom. The Fig. 4.2 shows the ambient camera motion in each shot class except for establishing shots where the camera remains stationary. Both aerial and bird-eye shots are captured from a high flying platform. The former class of shots have a strong perspective distortion, while the latter being taken from a camera ortho-normal to the ground plane, show affine transformation properties between consecutive frames. Crane or boom shots involve vertical motion of camera

which may include simultaneous movement along x or y axes A dolly shot, on the other hand, is taken by placing the camera on a platform that moves smoothly on ground without any movement along z-axis. Pan and tilt shots are associated with camera rotation along z and y-axes respectively. A zoom shot, does not involve any physical camera motion. It is characterized by the change in focal length, which is an internal camera parameter. All of these motions can be efficiently captured by the projective transformation model.

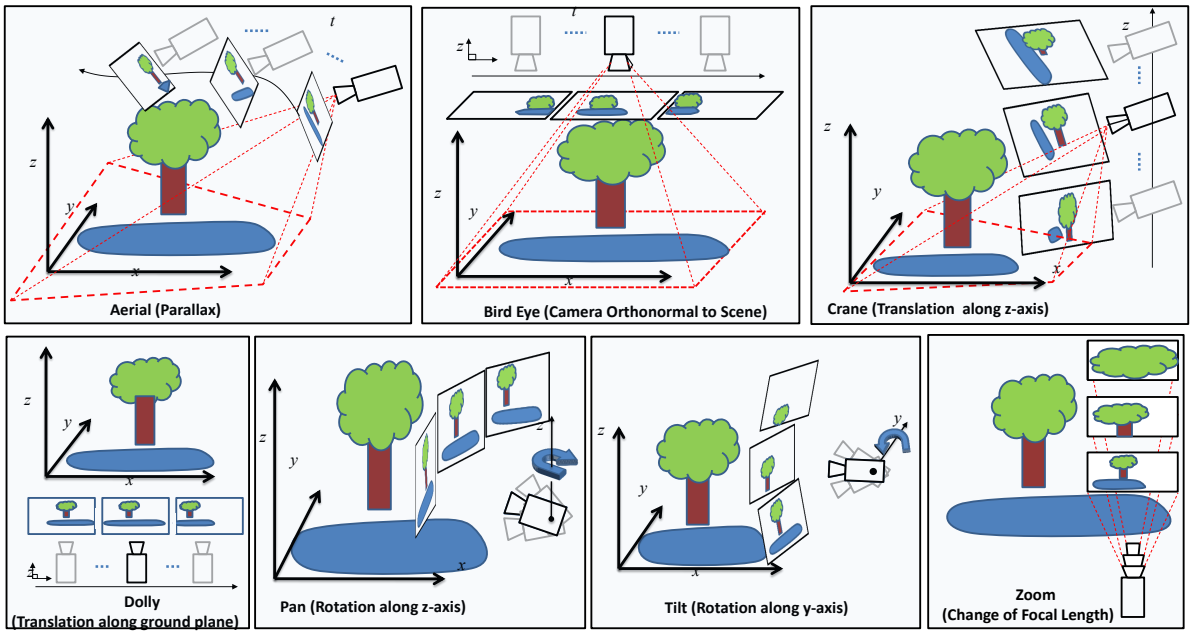


Figure 4.2: Schematic diagram showing different types of shots –**Top Row**: The first two figures show aerial and bird-eye shots. In both shots the camera is attached to a high flying platform and has its characteristic motion in 3D. In case of aerial shot, there is a strong perspective which is absent in case of bird-eye shots. The third figure shows a crane shot where the crane moves along z-axis with no simultaneous motion along x or y axis. Red lines show the field of views of each camera in a particular shot setting. **Bottom Row**: The first figure shows a dolly shot where the camera is on a platform that undergoes smooth translation along the ground plane. The next three figures show pan, tilt and a zoom shot. Pan and tilt shots are associated with camera rotation along z-axis and y-axis respectively. A zoom shot as shown, does not involve any physical camera motion. The change of focal length in this case is indicated using dotted lines with different sized lenses.

4.3 Motion Parameter Extraction

We employ a feature based method to estimate homography between consecutive frames of a given shot. In our technique, SURF features [15] are detected on each pair of frames on a dense sampling basis. Correspondence between features are established using a nearest neighbor search.

Given two sets of corresponding points $\{(x_1, y_1), \ldots (x_n, y_n)\}$, and $\{(x'_1, y'_1), \ldots (x'_n, y'_n)\}$ a homography $H = \{h_{ij}\}$, is a 3×3, 8 degrees of freedom projective transformation that models the relationship between two points (x, y) and (x', y') in the following way:

$$x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + 1}, y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + 1}. \tag{4.1}$$

Using a set of N corresponding points, we can form the following linear system of equations:

$$[{a_{x_1}}^T, {a_{y_1}}^T, {a_{x_2}}^T, {a_{y_2}}^T, \ldots, {a_{x_N}}^T, {a_{y_N}}^T]^T \mathbf{H} = 0, \tag{4.2}$$

where $\mathbf{H}, a_x, a_y$ are the following vectors:

$$\begin{aligned} \mathbf{H} &= [h_{11}, h_{12}, h_{13}, h_{21}, h_{22}, h_{23}, h_{31}, h_{32}, h_{33}]^T, \\ a_x &= [-x_i, -y_i, -1, 0, 0, 0, x'_i x_i, x'_i y_i, x'_i]^T, \\ a_y &= [0, 0, 0, -x_i, -y_i, -1, y'_i x_i, y'_i y_i, y'_i]^T. \end{aligned} \tag{4.3}$$

Eqn.(4.2) is solved using random sampling consensus technique [49] that iteratively minimizes the back-projection error, defined as:

$$\sum_i (x'_i - x''_i)^2 + (y'_i - y''_i)^2 \tag{4.4}$$

where,

$$x''_i = \left(\frac{h_{11}x_i + h_{12}y_i + h_{13}}{h_{31}x_i + h_{32}y_i + h_{33}} \right), \tag{4.5}$$

and,

$$y_i'' = \left(\frac{h_{21}x_i + h_{22}y_i + h_{23}}{h_{31}x_i + h_{32}y_i + h_{33}} \right) \tag{4.6}$$

For practical purposes, the last element of the matrix h_{33} is normalized to 1 which gives 8 transformation parameters between each frame-pair. Except for h_{13} and h_{23}, which indicate translational motion along x and y axes respectively, these parameters are not individually meaningful (this is experimentally validated in Section 4.6). However, since they represent a transformation, they can be mapped efficiently to some subspace that preserves the internal structure of the transformation. We resort to Lie algebra for projective group to establish this mapping.

4.4 Lie Algebra Mapping of Projective Group

Recently, Lie algebra is made popular by the authors of [91, 93, 94] to solve a wide range of tasks in computer vision. The algebraic representation of affine and projective transforms facilitates the use of learning methods by providing an equivalent vector space that preserves the geometric transformation structure under linear operations.

Homographies belong to the projective group which has multiplicative structure. This group is neither closed under vector addition nor scalar multiplication, and therefore treating it as a linear space for classification results in undesirable effects. This is because nearest neighbor or SVM based classification do not consider geometric constraints which apply to projective groups since they belong to a nonlinear manifold. The Lie algebra mapping of the projective group is a 3×3 matrix in homogeneous space which relates to the homography matrix H through an exponential function as:

$$H = \exp(M) = I + \sum_{k=1}^{\infty} \frac{1}{k!} M^k, \tag{4.7}$$

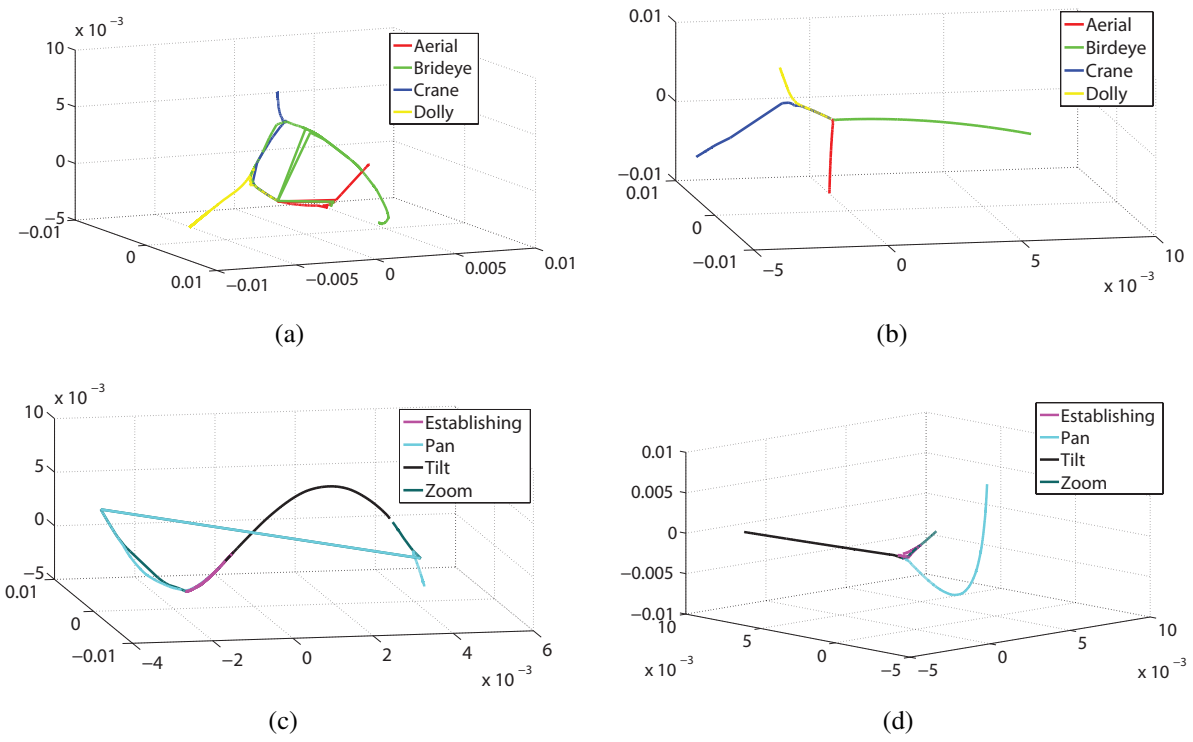


Figure 4.3: Lie Algebraic representation of homographies of typical shots: (a) shows the pure accumulative homographies before mapping in Lie algebra for aerial, bird-eye, crane, and dolly shots. Each shot is shown in trajectories of different colors. (b) signifies the effect of mapping (For visualization purposes, we reduce the 8-dimensional vectors into 3 dimensions). In (a), we observe a large overlap between the cumulative homographies from different shot classes. Due to the Lie algebra mapping, the trajectories from each shot class show fair amount of separation (less overlap). Similarly, (c) and its corresponding mapping in (d) show a similar trend for the remaining categories of shots:establishing, pan, tilt, and zoom. In this space when two points are close, it means they are similar in original space. We expect the trajectory of each shot starting from a point near the origin $(0, 0, 0)$ and diverging from the other classes as the accumulative homographies of different shot categories are different.

Alternatively,

$$M = \log(H) = \sum_{k=1}^{\infty} \frac{-1^{k+1}}{k} (H - I)^k. \tag{4.8}$$

Due to linearity in the Lie algebraic representation, M can be written as the linear combination of orthogonal bases as:

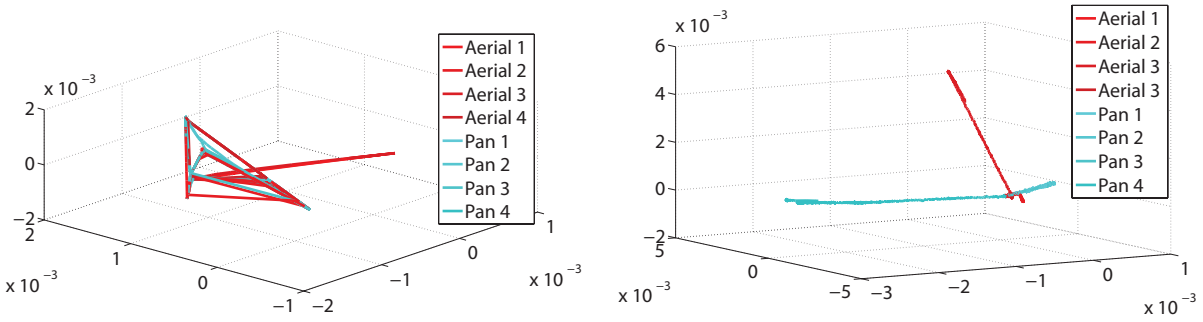
$$M = \sum_{i=1}^{8} \alpha_i G_i \tag{4.9}$$

where, G_i are also called generators of the Lie group [41]. It is shown in [41] that for infinitesimal

transformations near identity, the higher order terms in (4.8) can be ignored. Thus, α_i can be computed by projecting the first order approximation of M i.e. $H - I$ on G_i. In principle, as long as the bases are orthogonal, Eqn. 4.9 is valid. We select the following generators since they are already established in literature [41] and have injective mapping with the projective group of transformations:

$$G_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad G_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad G_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad G_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$
$$G_5 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad G_6 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad G_7 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix} \quad G_8 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} . \quad (4.10)$$

Using the above, the frame by frame homography matrix can be represented by $\{\alpha_i\}$ in an equivalent vector space. This can be easily illustrated with the help of the Fig. 4.3 and Fig. 4.4.



(a) (b)

Figure 4.4: Intra class similarity in Lie Space: four instances of aerial and pan shot classes are shown in different shades of red and cyan respectively. (a) shows accumulative homographies in the original 8-dimensional projective space. All the eight shots are intermingled and hence the similarity between the same category of shots is not visibly prominent. (b) shows clear separability between shot of the two classes and similarity between shots of same classes after the Lie Group mapping (For visualization purposes, we reduce the 8-dimensional vectors into 3 dimensions).

Fig. 4.3 visualizes various shots in their original space and their corresponding Lie algebra representations. In the mapped space, it can be observed that the trajectories representing different shots are clearly separable as compared to their original space. Fig. 4.4 demonstrates another important aspect of the mapping where we observe clear separation between shots of same classes which was not so obvious in the original projective space. Thus a shot can be represented by a sequence of these mappings, as 8 different time series.

4.5 Feature Extraction from Time Series

The different time series obtained after sequential arrangement of the Lie-group coefficients could be imagined as trajectories emanated from a nonlinear dynamical system (camera movement in 3-D space). It may be tempting to fit these trajectories into splines or simple models by finding the parameters that best explains the data, however classification using these models is complex and can be badly distorted by outliers as they usually do not have any structural interpretation. We perform extensive evaluations on the feature selection process that is necessary to describe our data. The remaining of this section gives some theoretical details on the feature computation. In our first method, we compute statistically invariant features from trajectories, not making any assumption of the nonlinear dynamical system that generates it. The second methods is inspired by the work [6] of Ali *et al.* wherein they introduced chaotic invariants of NLDS (3D trajectories of human body joints) for classification of human actions. Finally, motivated by the success of Hankel or Topelitz Matrices in detection of sequential signals [38, 87], we compute invariant features which can be used in the context of classification.

4.5.1 Statistically Invariant Features

We compute the following statistics from each dimension of the 8-dimensional trajectory separately. Let X be a single dimension from the trajectory. The separate statistical features

computed from X are as follows: mean(μ), variance(σ), first and last order statistics ($X_{(1)}, X_{(n)}$), range ($|X_{(1)} - X_{(n)}|$), average crossing rate ($n(\frac{d(X-\mu)}{dt})/n(X)$, dt being temporal interval, $n(.)$ is the cardinality function), average root mean square, mean and variance of skew ($\frac{(X-\mu)^3}{\sigma^3}$), signal entropy, mean and variance of kurtosis ($\frac{(X-\mu)^4}{\sigma^4}$). In addition, we compute 28 pairwise correlations between each of the eight dimensions of the trajectory. Finally the sum and the squared sum of all the dimensions is computed. This results in a total of $8 \times 12 + 28 + 2 = 126$ features.

4.5.2 Chaotic Invariant Features

Since, the above features are based on pure statistics, they do not exploit any characteristics of the underlying nonlinear dynamics of the system generating these trajectories. This motivates us to investigate analysis of these trajectories using dynamics of chaotic systems. Any such system is characterized by a set of metric, dynamical and topological organization of orbits that can be quantified using invariants of the system. These invariant features can be considered as signature of a particular system and hence can be used as features for final representation of a given shot.

Consider the multivariate time series $\boldsymbol{\alpha}_{\boldsymbol{i=0}}^{\boldsymbol{n}} = [\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, \ldots \boldsymbol{\alpha_n}]$ ($\boldsymbol{\alpha} \in R^8$), constructed from the Lie algebra sequential mapping of homography coefficients from a shot. In order to extract meaningful information from this chaotic system, we first need to embed it to a phase space. This is achieved by Takens theorem [157] which involves computation of delay parameter (τ) and optimal embedding dimension (m). Given the embedding dimensions and the delay parameters,

the phase space embedding can be defined as:

$$Y = \begin{pmatrix} \alpha_0^{(1)} & \alpha_\tau^{(1)} & \dots & \alpha_{(m-1)\tau}^{(1)} \\ \alpha_0^{(2)} & \alpha_\tau^{(2)} & \dots & \alpha_{(m-1)\tau}^{(2)} \\ \alpha_0^{(3)} & \alpha_\tau^{(3)} & \dots & \alpha_{(m-1)\tau}^{(3)} \\ \alpha_0^{(4)} & \alpha_\tau^{(4)} & \dots & \alpha_{(m-1)\tau}^{(4)} \\ \alpha_0^{(5)} & \alpha_\tau^{(5)} & \dots & \alpha_{(m-1)\tau}^{(5)} \\ \alpha_0^{(6)} & \alpha_\tau^{(6)} & \dots & \alpha_{(m-1)\tau}^{(6)} \\ \alpha_0^{(7)} & \alpha_\tau^{(7)} & \dots & \alpha_{(m-1)\tau}^{(7)} \\ \alpha_0^{(8)} & \alpha_\tau^{(8)} & \dots & \alpha_{(m-1)\tau}^{(8)} \end{pmatrix} \tag{4.11}$$

where the numbers in parentheses represent each dimension of the 8-dimensional α.

Any chaotic system is characterized by a set of metric, dynamical and topological organization of orbits that can be quantified using invariants of the system. These invariant features can be considered as signature of a particular system and hence can be used as features for classification task. We use the same technique applied in [6] to determine the following chaotic invariants:

Maximal Lyapunov Exponent is a quantity that characterizes the rate of divergence of infinitesimally close trajectories in phase space. Assume an arbitrary point $p(i)$ in phase space surrounded by a set of points $p(k)$ within distance ϵ. The average distance between a reference trajectory emanating from $p(i)$ to all trajectories emanating from $p(k)$ can thus be computed as a function of relative time Δn using:

$$D_i(\Delta n) = \frac{1}{r} \sum_{s=1}^{r} |\alpha_{k+(m-1)\tau+\Delta n} - \alpha_{i+(m-1)\tau+\Delta n}| \tag{4.12}$$

where r is the total number of points. From Eqn.(4.12), we compute $S(\Delta n) = \frac{1}{c} \sum_{i=1}^{c} \log(D_i(\Delta n))$, c being the typical number of sampled points for which the above process is repeated. The Maximal Lyapunov Exponent is calculated as the largest value of the slope of $S(\Delta n)$ against different

values of Δn.

Correlation Integral quantifies the density of points in phase space by performing a normalized count of pairs of points lying within a radius ϵ. Mathematically it is defined as:

$$C(\epsilon) = \frac{2}{N(N-1)} \sum_{t=1}^{N} \sum_{s=t+1}^{N} \phi(\epsilon - ||p(t) - p(s)||) \tag{4.13}$$

where $\phi(.)$ is a Heaviside function.

Correlation Dimension characterizes the dimensionality of the phase space occupied by a set of random points. It measures the rate of change in the density of phase space with respect to a neighborhood (ϵ) and is calculated as:

$$D_c = \lim_{\epsilon, \epsilon' \to 0^+} \frac{\log\left(\frac{C(\epsilon)}{C(\epsilon')}\right)}{\frac{\log(\epsilon)}{\log(\epsilon')}} \tag{4.14}$$

These chaotic invariants can be used as input feature vector to any classifier. We explored two different types of classification strategies in our experiments which is covered in section 4.6.

4.5.3 Hankel Matrix based features

Given a finite sequence of coefficient vectors length n ($\boldsymbol{\alpha}^{(0)} \ldots \boldsymbol{\alpha}^{n}$) the Hankel matrix can be can be constructed as follows, whose entries are the same along the anti-diagonals:

$$K_i = \begin{pmatrix} \boldsymbol{\alpha}_i^{(0)} & \boldsymbol{\alpha}_i^{(1)} & \boldsymbol{\alpha}_i^{(2)} & \ldots & \boldsymbol{\alpha}_i^{(n-r+1)} \\ \boldsymbol{\alpha}_i^{(1)} & \boldsymbol{\alpha}_i^{(2)} & \boldsymbol{\alpha}_i^{(3)} & \ldots & \boldsymbol{\alpha}_i^{(n-r+2)} \\ \boldsymbol{\alpha}_i^{(2)} & \boldsymbol{\alpha}_i^{(3)} & \boldsymbol{\alpha}_i^{(4)} & \ldots & \boldsymbol{\alpha}_i^{(n-r+3)} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{\alpha}_i^{(r-1)} & \boldsymbol{\alpha}_i^{(r)} & \boldsymbol{\alpha}_i^{(r+1)} & \ldots & \boldsymbol{\alpha}_i^{(n)} \end{pmatrix}, \tag{4.15}$$

where r is an integral estimate on the number of entries of the j-th column vector that are sufficient to express the subsequent $(j+1)$-th column in K_i. It is previously shown in [38, 87] that matrices of the above form capture the dynamical structure of a system in a meaningful manner, which can be characterized by the orthogonal basis of the above matrix. The orthogonal basis is obtained after performing Singular Value Decomposition on $K_i K_i^T$ as follows:

$$[U \quad \Sigma \quad U^T] = \text{SVD}(K_i K_i^T), \tag{4.16}$$

where U is the orthogonal matrix containing the eigen vectors and Σ contains the corresponding eigen values in a diagonal matrix. All Hankel matrices are normalized using the Frobenius norm for matrices using the following equation:

$$\hat{K}_i = \frac{K_i}{\text{trace}(K_i K_i^T)^{\frac{1}{2}}}, \tag{4.17}$$

In order to compute a uniform length descriptor for final classification using a linear SVM, we reproject $K_i K_i^T$ on to its largest eigen vector as follows:

$$\mathbf{v}_i = \mathbf{u_1}^T (K_i K_i^T), \tag{4.18}$$

$\mathbf{u_1}^T$, being the largest eigen vector. This results in a r dimensional descriptor for a shot which in all our experiments are fixed to 8 to maintain sufficient overlap between column vectors of the matrix.

All the above features can be used separately for classification. In the next sections we share some details on the experimental protocol we follow. At first, we discuss our dataset of 8 distinct category of shot classes based on cinematographic guidelines. The following section provides implementation specific details on the various stages involved in our computational workflow. This is followed by results and discussion. On a separate note, we describe how this shot classification

technique can be integrated into large scale complex event recognition, backing our claim with results.

4.6 Experiments on Cinematographic Shot Dataset

4.6.1 Dataset of Cinematographic Shots

Most of the earlier chapters [44, 126, 153, 195] on this topic evaluate their respective approaches on their own private collections, which are not made available. In order to contribute to the research community, we make an attempt to build the first dataset of this kind which is reusable, expandable and publicly available. Our dataset consists a clean and an unconstrained part. The clean part has videos downloaded from high resolution, professional stock video [1] while the unconstrained part contains videos from amateur consumer uploaded videos found in YouTube. The unconstrained part contains videos uploaded by amateur users that typically have fair amount jitters caused due to unstable mounts. These two separate sources were used for two different experiments to validate the efficiency of our shot representation. Each videos in the dataset conforms to either one of eight categories, namely: (1) Aerial, (2) Bird eye, (3) Crane, (4) Dolly, (5) Establishing, (6) Pan, (7) Tilt, and (8) Zoom. Each video is carefully screened by 3 human observers with good cinematographic knowledge to ensure there is no mixing up of camera motions in a particular video. Note that this is a difficult task since most shots do not occur in isolation as pointed out in [148]. Finally all videos are resized to an approximate resolution of 480×360 keeping the aspect ratio locked. Some sample frames from the clean part of our dataset are shown in Fig. 4.5. Table 4.1 contains some statistics of our dataset.

[1] http://www.gettyimages.com

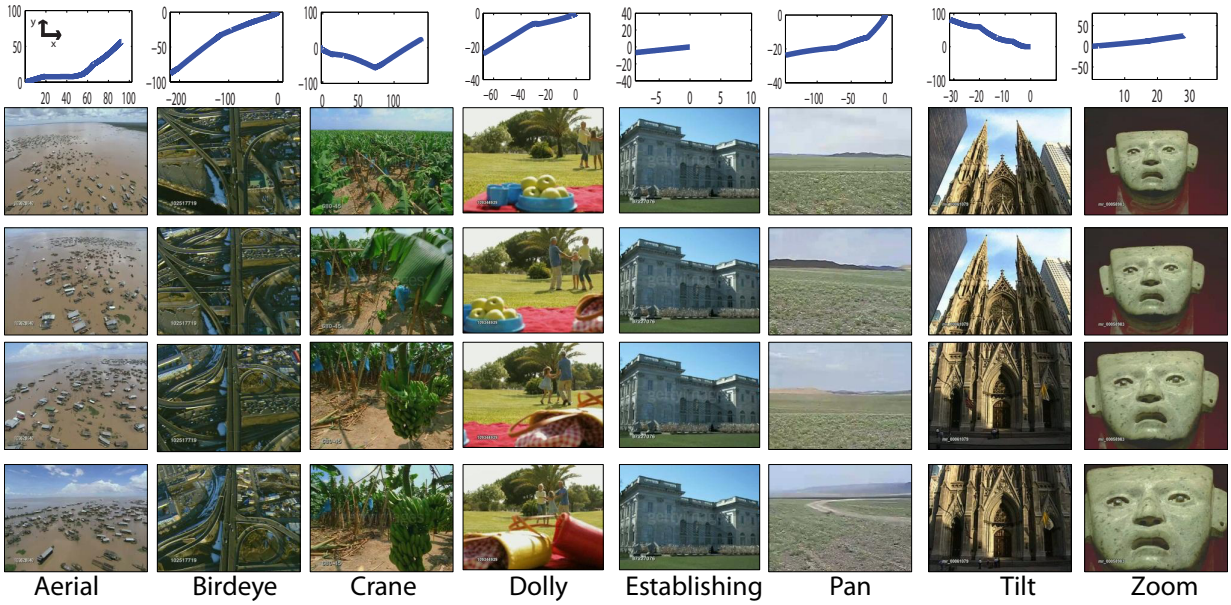


Figure 4.5: Cinematographic shot dataset: Each column in the figure represents a typical shot category. The top row shows the trajectory against x and y axes of the image plane (obtained by tracking points). For establishing and zoom shots we can observe that there is very limited motion. The second row contains the initial frame from the shot. Subsequent rows show samples 50 frames apart. Images from top to bottom provide an idea of the camera motion as the shot progresses.

Table 4.1: Some statistics from our cinematographic shot dataset (Unc. stands for the unconstrained part of the dataset).

Shot Category	Examples		Total # of Frames	
	Clean	**Unc.**	**Clean**	**Unc.**
Aerial	30	10	18122	3622
Bird-eye	30	7	18644	1578
Crane	43	8	20304	1226
Dolly	32	8	22241	1185
Establishing	36	9	20454	1256
Pan	30	7	22954	1806
Tilt	31	7	12718	3998
Zoom	31	8	14876	2320

4.6.2 Experimental Setup

We use an OpenCV based implementation of the SURF extraction [15] and use an approximate nearest neighbor search algorithm [110] to obtain point correspondences which is later required for homography estimation. The normalized homographies and their corresponding Lie-algebra mappings are used in a bag of words framework [148] under different codebook configurations in the range : $128, 256, \ldots, 2048$ and these help us investigate the efficacy of our shot representation incrementally. These two are abbreviated as BoHM (Bag of Homographies) and BoLC (Bag of Lie Algebra representations of homographies) in Table 4.2. In this setting, SVMs with histogram intersection kernel is used for classification using a 10 fold cross validation scheme.

For the time series constructed after stacking corresponding Lie grooup coefficients, we independently evaluate three sets of features: STAT (Statistically Invariant), CI (Chaotic Invariant), and HNK (Descriptors from Hankel Matrix). In order to evaluate how our method performs against a more accurate camera trajectory estimation technique (using full structure from motion [150]), we compute statistically invariant features on top of 3-D camera trajectories (abberviated as TFT). Trajectories are obtained after connecting 3-D camera locations in space, temporally based on their frame indices. Although features extracted using this method are very discriminative, the trajectory computation in itself a computationally prohibitive task as the 3D reconstruction algorithm needs an exhaustive set of points datapoints from all frames in a video to solve a complex optimization problem. This makes this technique a misfit for large-scale internet videos.

Finally, we compare our proposed method with our implementation of two relevant algorithms: MSL (Motion Slices [118]) and HF (Threshold selection on fundamental matrices [193]). The former represents a shot using tensor histogram of spatio-temporal slices while the latter uses a combination of homography and fundamental matrix to represent a shot.

4.6.3 *Results and Discussions*

In this section, we provide detailed experimental evaluation of our proposed method on our shot dataset. In Fig. 4.6 we show the effect of temporal sampling rates for homography estimation on the overall classification performance. As discussed previously, the sampling rates can be perceived as the number of frames that are skipped between any given pair of frames before computing the homography between that pair. Typically, the larger the gap between two sampled frames, the more the homographies deviate from identity as the relative inter-frame motion increases. The average accuracy reaches its peak when the sampling interval is 4, i.e. homography is computed between pairs separated by four frames. This can be explained with the help of evidence from homography computation which is primarily noisy for smaller temporal intervals. At interval lengths larger than 4, the homography violates the primary assumption for Lie group mapping which states that the transformation should be approximately equal to identity.

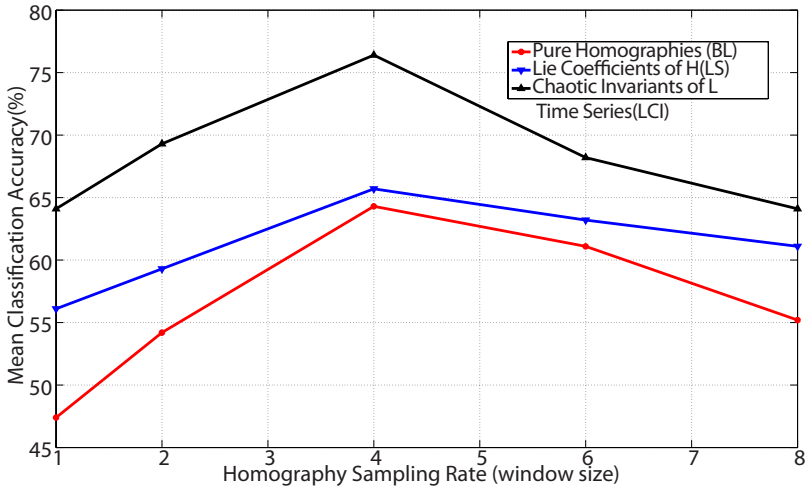


Figure 4.6: The above Figure shows effect of different temporal sampling intervals on homography computation. Red: pure homographies, Blue: Lie-coefficients of homography elements, Black: Time series based features.The average classification performance reaches its peak when the sampling interval is 4, i.e. homography is computed between pairs separated by four frames.

In order to show how the time series based features affect the final classification, we conducted a bag-of-features based experiments on different codebook sizes. The maximum avg. ac-

curacy is observed with a codebook size of 1024 and the results are reported in the second and third columns of Table 4.2. This is in accordance with our hypothesis that temporal information captured from time series is more discriminative than an orderless bag-of-features representation, as seen in the forth and fifth columns. The statistical features perform slightly worse as compared to the chaotic invariant features which capture the nonlinear dynamics of the trajectories. The last two columns show accuracies of our implementation of two state-of-the-art methods.

Table 4.2: Quantitative comparison of our proposed shot representation method against other methods (SVM is used as classifier for all cases with 10-fold cross validation): Rows show the average accuracy on individual classes, Column 2: Naive Bag-of-feature representation on Homographies (BoHM) Column 3,4: MSL [118] and HF [193] are our implementations of two state-of-the-art methods. It should be noted that these two approaches do not apply to all the kind of shots in the dataset. Column 5: Bag of words representations on Lie group mappings of homographies with a codebook size of 1024 (BoHM, BoLC). Column 6–8: Proposed shot representation with Statistically invariant (STAT), Chaotic invariant (LCI) and Hankel Matrix (HKL) based invariant features computed from time-series of Lie coefficients. Column 7: Statistically invariant features computed on 3D trajectories (TFT) estimated from the shots using [150], this is analogous to ground truth.

Category	BoHM	HF [193]	MSL [118]	BoLC	STAT	LCI	HKL	TFT
Aerial	21.4 ± 1.2	–	–	30.0 ± 1.1	86.2 ± 0.4	88.2 ± 0.4	90.1 ± 0.8	91.4 ± 0.7
Bird-eye	54.8 ± 0.9	–	20.1	54.8 ± 1.0	89.3 ± 0.6	89.4 ± 0.2	89.4 ± 0.5	92.1 ± 0.5
Crane	44.2 ± 0.6	–	–	60.5 ± 0.9	73.3 ± 0.3	71.5 ± 0.7	74.2 ± 0.6	75.1 ± 0.6
Dolly	43.8 ± 0.8	–	–	65.6 ± 0.8	62.0 ± 0.5	65.9 ± 0.8	65.4 ± 0.9	65.6 ± 0.4
Establishing	86.1 ± 0.9	100.0	–	94.4 ± 0.4	99.1 ± 0.4	96.8 ± 0.6	97.0 ± 0.1	96.7 ± 0.4
Pan	83.3 ± 1.1	65.3	3.3	93.3 ± 0.2	63.5 ± 0.6	66.1 ± 0.8	68.1 ± 0.4	69.1 ± 0.6
Tilt	66.7 ± 1.2	55.2	26.6	70.0 ± 0.4	76.7 ± 0.5	79.2 ± 0.5	81.0 ± 1.1	81.1 ± 0.9
Zoom	51.6 ± 1.6	22.7	45.1	51.6 ± 0.7	57.1 ± 0.8	59.2 ± 0.9	58.2 ± 1.0	61.1 ± 1.2
Avg.	56.5 ± 0.5	60.8	23.7	65.0 ± 0.7	75.9 ± 0.5	77.0 ± 0.6	77.9 ± 0.7	79.0 ± 0.6

Finally, we show how a good shot classification model learned on the clean portion of our shot dataset could be applied to perform shot classification on unconstrained videos from YouTube. Fig. 4.7 shows the confusion in classification across shots from the clean dataset. Some confusion is evident between aerial and bird-eye classes because of their obvious similarity. Fig. 4.8 shows the confusion across different classes when we use the classification model learned on the clean portion of the dataset. There is a significant drop in performance, however we conjecture that this is mainly due to the large variation between training and testing data.

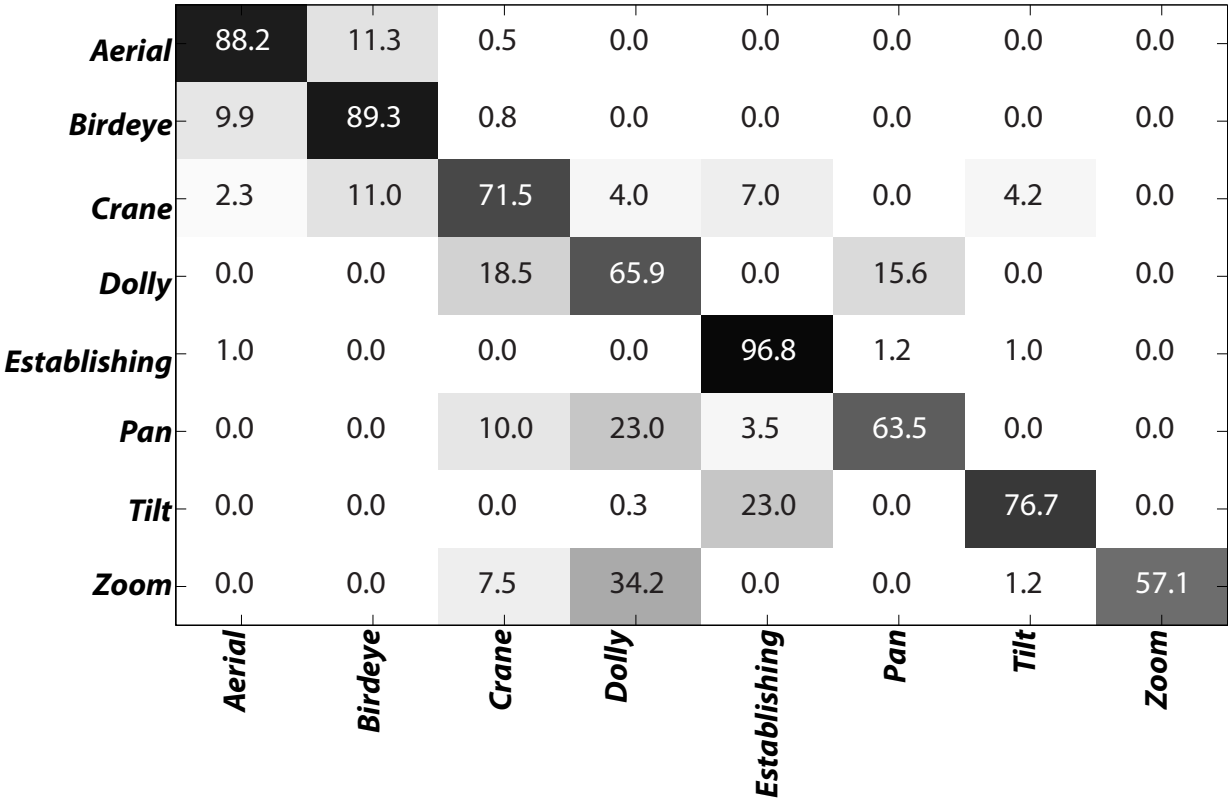


Figure 4.7: The above figure shows confusion across different examples from the clean part of the dataset.

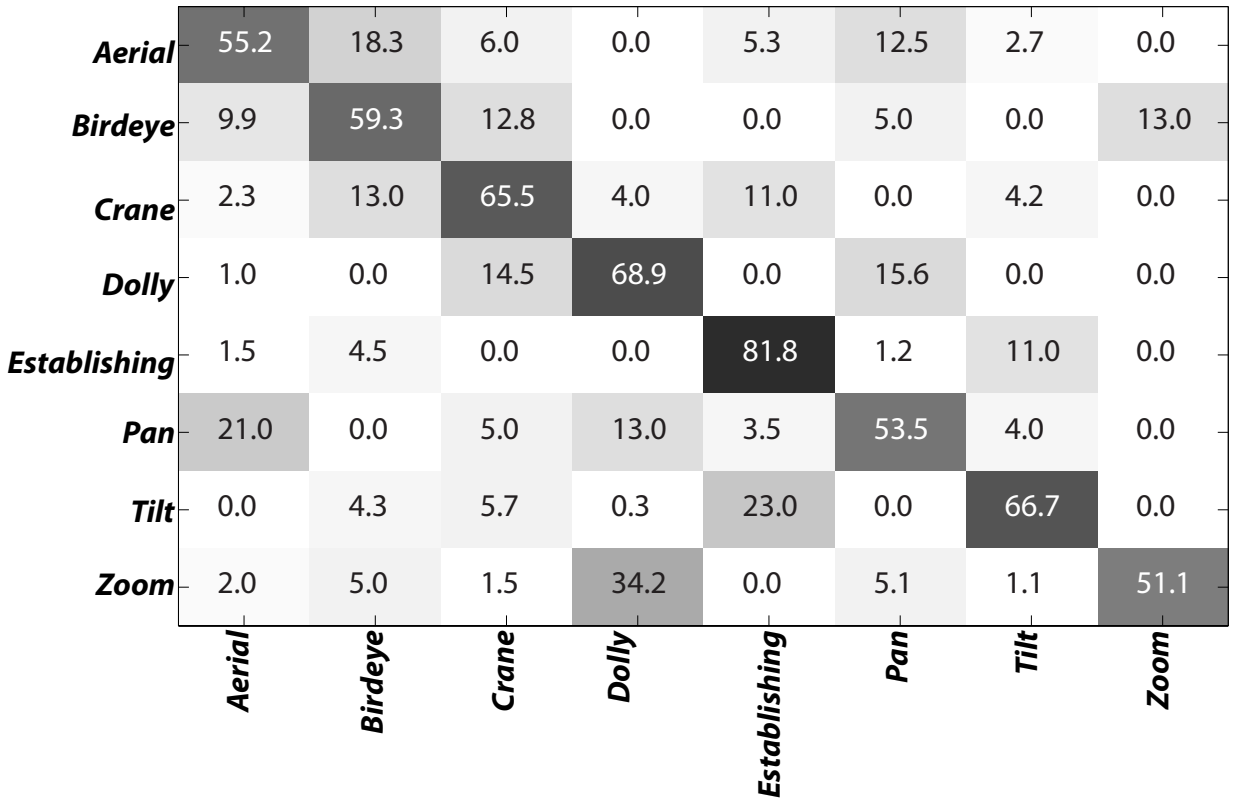


Figure 4.8: This figure shows the confusion matrix from unconstrained shots. In this case we do not perform any training, we directly use the learned model from the clean portion to predict shot labels in the unconstrained data.

Table 4.3 shows the typical computational aspect of different computational steps involved in the entire computational workflow. An asymptotic analysis of all the algorithms discussed here is out of scope of this chapter.

With this, we conclude our experiments on the cinematographic shot dataset and move on to a more challenging task: Classifying events captured in unconstrained YouTube-like videos based on their respective ambient camera motion.

Table 4.3: **Computational Aspects**: Each row indicates a computational step, implemented in C++/OpenCV. From top to bottom: Feature Extraction (FE), Homography Estimation (HE), Vector Space Mapping (VSM), Time Series Feature Computation (TSFC). The speed is recorded for a 320×240 video containing 300 frames on a standard desktop hosting a 2.4 Ghz CPU.

Step	Speed (in ms)	Size Dependence	Parallelizable
FE	5300	Yes	Yes
HE	75	Yes	No
VSM	8	N/A	No
TSFC-STAT	3	N/A	No
TSFC-LCI	12	N/A	No
TSFC-HKL	4	N/A	No

4.7 Recognition of Complex Events using Camera Motion

Recently, NIST has released the Multimedia event detection competition [2] dataset which consists of videos from 15 event categories namely (1) Attempting a board trick, (2) Feeding an animal, (3) Landing a fish, (4) Wedding ceremony, (5) Working on a woodworking project, (6) Birthday party, (7) Changing a vehicle tire (8) Flash mob gathering, (9) Getting a vehicle unstuck, (10) Grooming an animal, (11) Making a sandwich, (12) Parade, (13) Parkour, (14) Repairing an appliance, and (15) Working on a sewing project. We use a subset of this dataset that has 2062 videos from all these 15 event categories for our experiments. Events like "Attempting a board trick" and "Parkour" usually have a lot of jittery camera motion coupled with pan and tilt motions. Similarly, videos depicting events such as "Wedding Ceremony" and "Birthday Party" are mostly captured by stationary cameras with limited pan and some amount of zoom. The goal of this experiment is to find out if we can leverage our proposed representation to capture these meaningful statistics from these amateur videos and perform crude event detection without resorting to any content extraction techniques.

Most of the videos in the TRECVID MED 2011 corpus are observed to have shots of the following kinds: pan, tilt, zoom, establishing. We divide each video into non-overlapping, fixed

[2] http://www.nist.gov/itl/iad/mig/med11.cfm

length shots of 100 frames and extract Hankel matrix based features from time-series constructed using the method suggested in Section 4.5. We perform shot detection using the classifiers trained on cinematographic shots dataset. The maximum detection probability subjected to a threshold (0.6), returned by any classifier is chosen as the shot label. For detections under the desired threshold, the shot is labeled unidentified. Thus for a given video, a sequence of detected shots is constructed with labels E (Establishing), P (Pan), T (Tilt), Z (Zoom) and U (Unidentified). Shot sequences from all 2062 videos are computed using the above methodology.

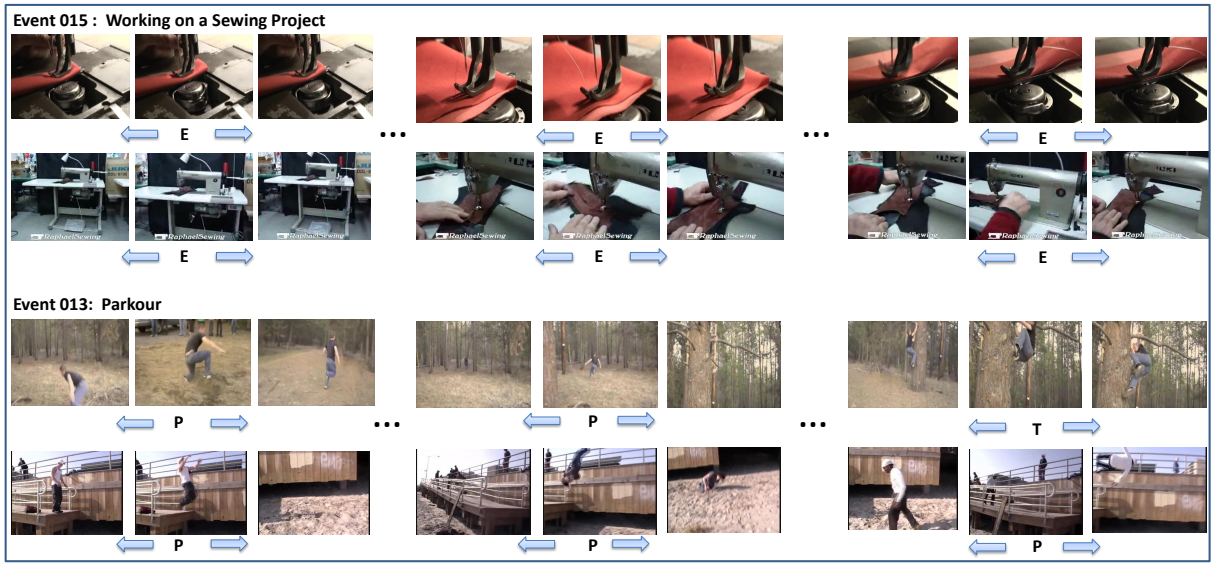


Figure 4.9: Camera Motion based Representation of Events: Top two rows represents two different videos each from an event class. Each video is divided into fixed length shots of 100 frames. Outputs from 4 shot classifiers: Establishing (E), Pan (P), Tilt (T), and Zoom (Z) shots are indicated under each shot..

Next, we train discrete Hidden Markov Models corresponding to each of the 15 different events with $50 - 50$ split of the data from each event class. For all HMMs, a first-order left-right topology is assumed with 10 states. Validation is performed against the remaining half of the videos from each event class. Figure 4.9 discusses this step where every row represents 3 shots from a video and their corresponding labels as detected by the classifiers. The labels are temporally concatenated and used as inputs for training event specific HMMs.

We show the performance of HMM based event detectors using Detection-Error Trade-off curves for top 5 events detected using the above methodology in Fig. 4.10. In Fig. 4.11, we compare this performance achieved after using classifiers trained on content based features (Bag-of-SIFT features + SVM). The DET curves give a better understanding of miss-detection against false alarm rates and is usually recommended for detection problems which was the primary focus of this experiment. We fix an reasonable operating region (6% false alarm with 75% mis detection) and measure the area under each curve intersecting this operating region. This is our single statistic (PAUC: Partial Area Under Curve) to compare the performance across all events. It is evident that our representation is highly favorable for events with discriminative camera motion (Attempting a Board trick, Changing a Vehicle tire, Birthday Party, Parade and Working on a Sewing project). It is interesting to note that for all these 5 events, our representation is as powerful as a content based representation. We confirmed that our representation of a video is complementary to existing content-based bag-of-feature representation. Our shot representation when used alongwith Bag-of-SIFT features in an early fusion framework, the overall event detection accuracy increases by 7%.

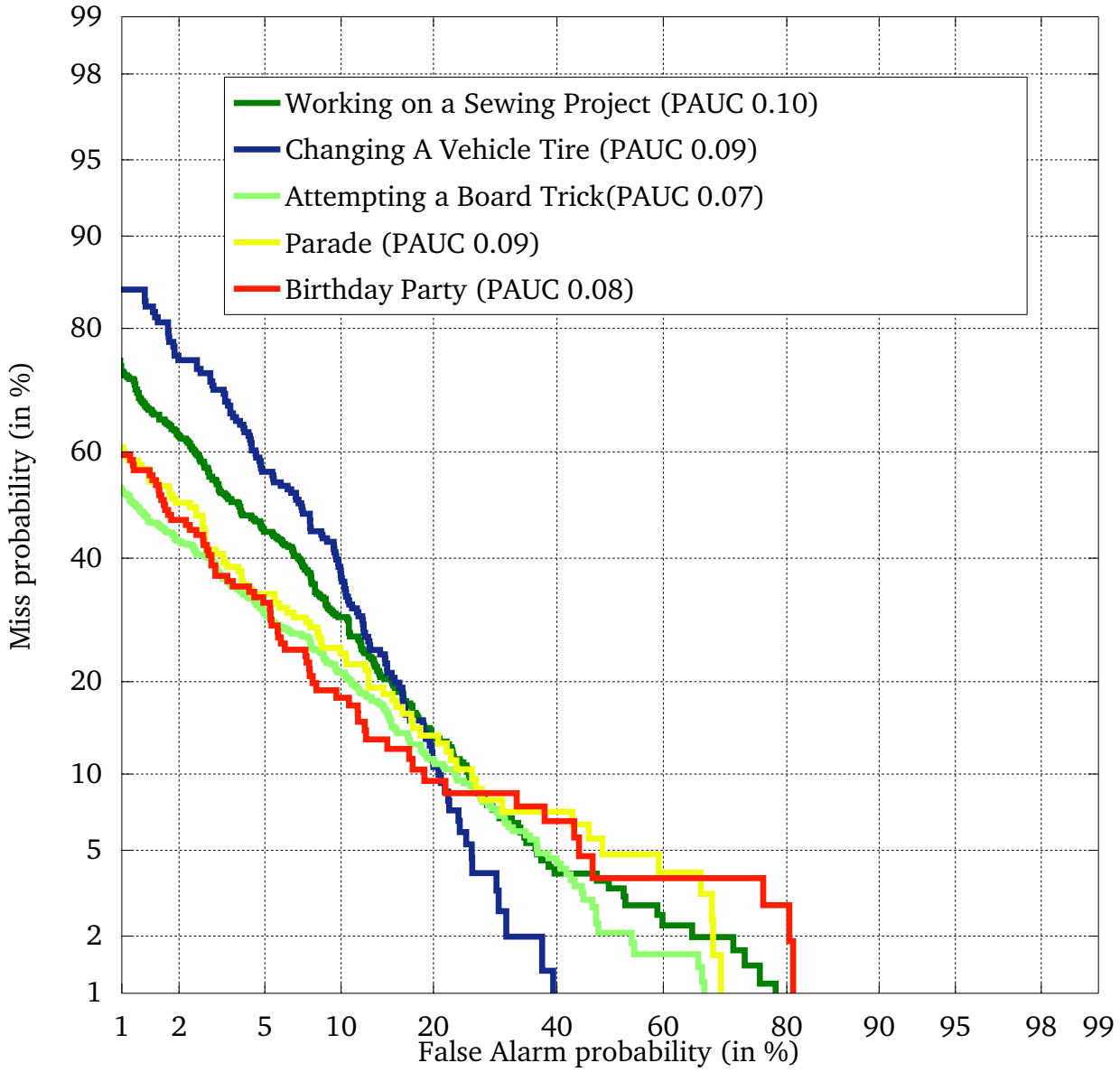


Figure 4.10: Detection-Error Trade off (DET) curves for 5 event classes best represented using our camera motion based features. .

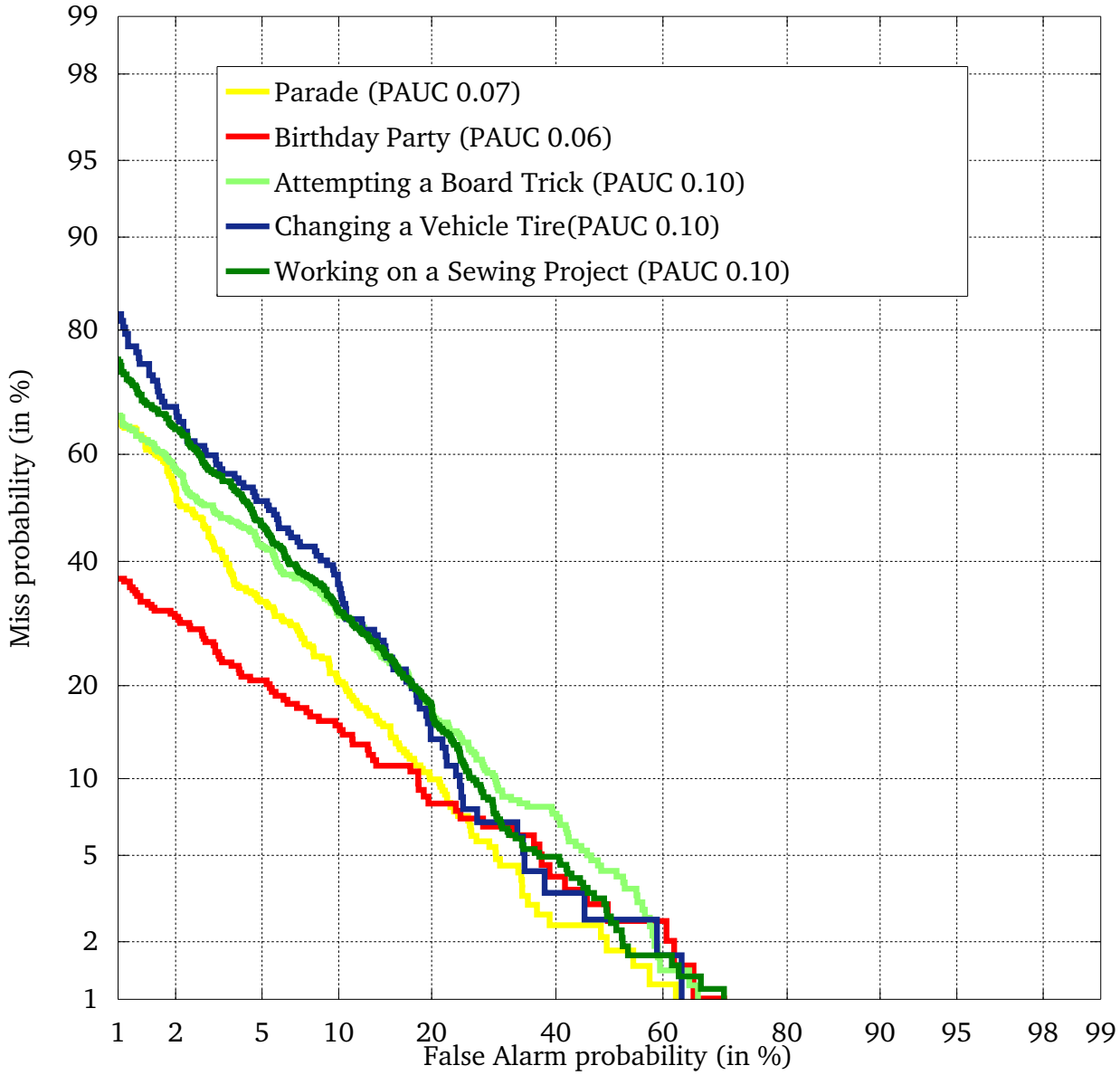


Figure 4.11: Detection-Error Trade off (DET) curves for corresponding event classes obtained using a content based feature representation (Bag-of-SIFT-features). .

4.8 Summary

We presented a novel set of methodologies [19] to perform robust shot classification based on camera motion adhering to cinematographic principles. In our approach, we first extracted camera motion from shots by computing frame to frame homographies. In order to represent homographies in a manageable space, we proposed the use of Lie algebra to obtain one to one linear mappings of the homographies. In order to exploit the temporal order these mappings, we compute features from time series constructed from these mappings. Our approach performs significantly better than the state of the art methods. As part of this work, we also introduced a cinematographic shot dataset that can be used by the research community to explore different avenues in this direction. Finally, we demonstrated the applicability of our proposed method to represent ambient camera motion in videos to develop insights towards solving a more challenging event detection problem. As part of future work, we intend to augment our complex event recognition framework with proper camera motion boundary detection [116, 117], instead of these fixed length segments.

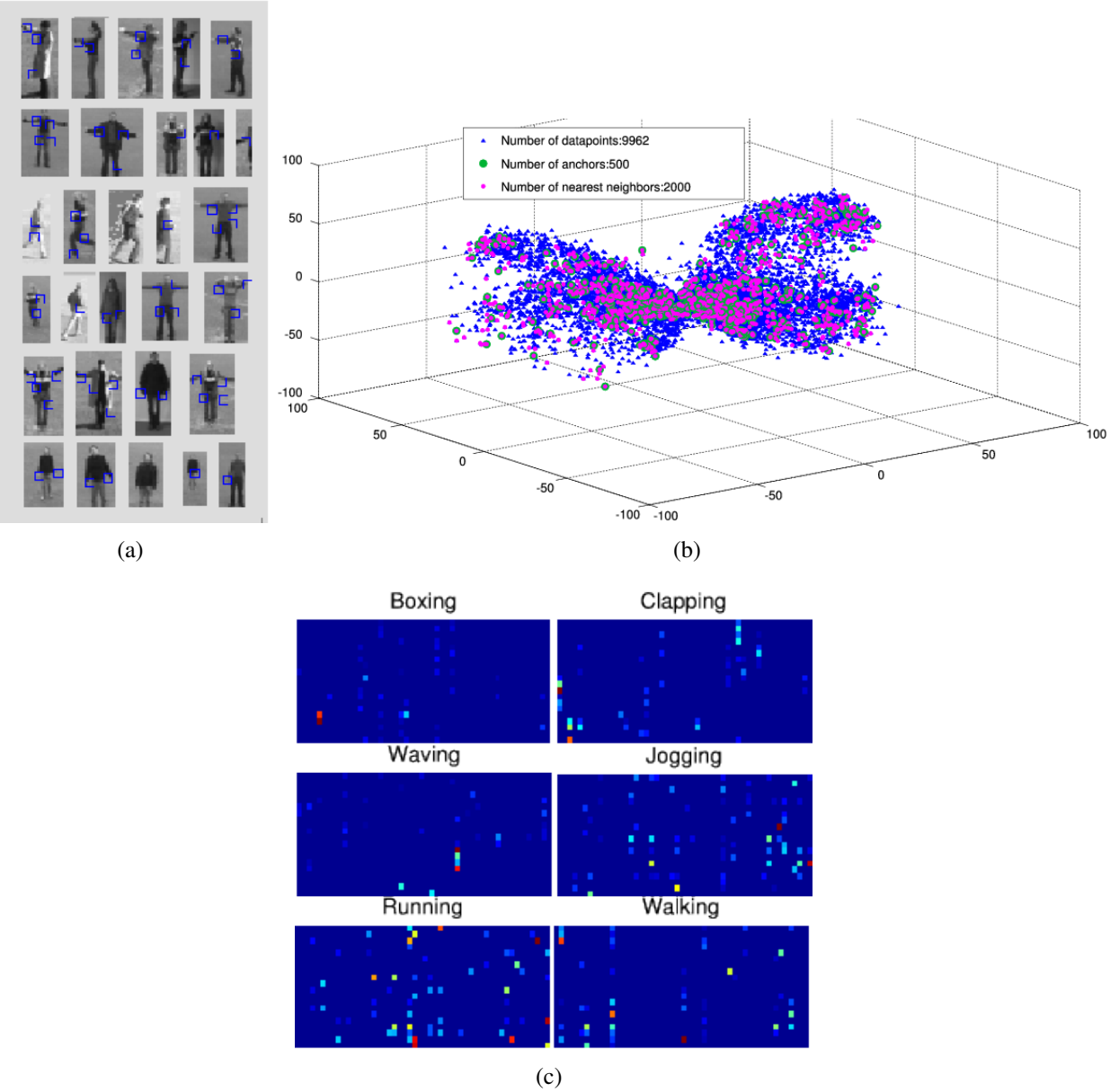
CHAPTER 5: PROBABILISTIC REPRESENTATION FOR EFFICIENT LARGE SCALE VISUAL RECOGNITION TASKS

5.1 Introduction

Automatic visual classification for content-based semantic interpretation of images and video remains an active area of research in computer vision. Canonical examples of such tasks include distinguishing an image of an urban scene containing buildings and street lights from that of a natural scene containing mountains, or detecting a particular type of human action (like running) observed in a video. Earlier approaches [36, 86, 147] have demonstrated the utility of constructing representations based on local features in images [101] and video [39, 81], analogous to words in text documents, enabling researchers to apply algorithms from text retrieval and classification to computer vision. These methods, popularly termed as "bag-of-words" (BoW) algorithms, advocate the creation of a vocabulary based on a clustering of visual words extracted from a corpus of images. A new image can then be expressed as a histogram (bag) of words using the designated vocabulary, thereby rendering it suitable for categorization using a classifier such as an SVM.

Our primary aim in this chapter is to propose a universal representation for images and videos that is based on sound statistical principles (maximum likelihood estimate of observed visual words in the given image). It inherits the benefits of soft-assignment [167] and is made computationally efficient through the use of bounded-support kernels and sampling-based (rather than clustering-based) anchor generation. Importantly, our representation [21] is completely compatible with existing classifier machinery used in bag-of-visual words approaches, enabling it to be easily integrated into existing real-world image and video recognition systems. Our experiments show the broad applicability of our representation to both image and video domains; wherever possible, we follow existing experimental methodology and avoid the temptation of tuning parameters to maximize performance on the dataset. Thus, our contribution is that of a novel representation

rather than the development of a complete system for either scene or action recognition.



(a) (b)

(c)

Figure 5.1: An illustration of the proposed representation [21] using the KTH human action dataset as an example. (a) Low-level feature extraction: Spatio-temporal features are extracted from input video sequences. (b) N anchors (shown as red triangular markers) are selected from the set of all video words (shown as blue triangular markers). Each feature contributes to nearby anchors (shown as green spheres), but in a manner that maximizes likelihood over the entire video.(c) A horizontally truncated sparse matrix 20×100 (originally 20×5000) corresponding to each of 20 training instances from each of the 6 classes is shown.

5.2 Approach

Let D_i be the set of visual features extracted from the i-th image I_i in a large collection of M labeled images $I \equiv \{I_1, I_2, \ldots, I_i, \ldots, I_M\}$. Thus, each D_i could be interepreted as a set of m-dimensional feature vectors whose cardinality may vary from image to image depending on the number of features extracted per image. Let us also denote by D the collection of all features extracted from all labeled training samples (I).

Consider a universal vocabulary of N representative visual features ($\{\mathbf{C}_j\}_{j=1}^{N}$), termed *anchors*. These anchors could be generated using traditional clustering or (as we suggest) sampled directly from D. Our proposed model assumes that visual features are generated i.i.d. from some distribution specified by a set of image-level parameters. Thus, we can express the probability of observing a particular feature $\mathbf{d}$ given an image I_i as:

$$p(\mathbf{d}|I_i) = \sum_{j=1}^{N} w_j K(\mathbf{d}, \mathbf{C}_j), \tag{5.1}$$

where $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_N)$ are the image-level parameters (weights) that control a kernel density function with kernel $K(.,.)$. In the proposed formulation, these weights $\mathbf{w}$ serve as the image representation and estimating them from the observed features is the primary task.

We propose determining $\mathbf{w}$ using a maximum likelihood estimator:

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w} \in \Delta} L(I_i, \mathbf{w}), \tag{5.2}$$

where Δ denotes all possible probability distributions for $\mathbf{w}$ and

$$L(I_i, \mathbf{w}) = \sum_{p=1}^{k} \log \sum_{j=1}^{N} w_j K(\mathbf{d}_p, \mathbf{C}_j).$$

k is the number of features extracted from image I_i. Eqn. (5.2) being a convex optimization

problem, has solutions that are globally optimal.

We propose the following computationally efficient iterative approach based on bound optimization that converges to the maximum likelihood representation. Let $\mathbf{w}'$ be the solution to Eqn. (5.2) at the current step and $\mathbf{w}$ be the solution at the next step, then $L(I_i, \mathbf{w}) - L(I_i, \mathbf{w}')$ is bounded as:

$$L(I_i, \mathbf{w}) - L(I_i, \mathbf{w}') = \sum_{p=1}^{k} \log \left[\frac{\sum_{j=1}^{N} w_p K(\mathbf{d}_p, \mathbf{C}_j)}{\sum_{j=1}^{N} w'_p K(\mathbf{d}_p, \mathbf{C}_j)} \right]$$

$$\geq \sum_{p=1}^{k} \sum_{j=1}^{N} \frac{w'_j K(\mathbf{d}_p, \mathbf{C}_j)}{\sum_{l=1}^{N} w'_l K(\mathbf{d}_p, \mathbf{C}_l)} \log \frac{w_j}{w'_j}. \tag{5.3}$$

The above bound can be easily verified by using Jensen's inequality for convex functions. $\mathbf{w}$ can be updated in each iteration using:

$$w_j = \frac{1}{Z} \sum_{p=1}^{k} \frac{w'_j K(\mathbf{d}_p, \mathbf{C}_j)}{\sum_{l=1}^{N} w'_l K(\mathbf{d}_p, \mathbf{C}_l)}, \tag{5.4}$$

where Z is a normalization term that guarantees $\sum_{j=1}^{N} w_j = 1$. Note that an approximation to Eqn. (5.4) can be obtained by initializing each of the elements of $\mathbf{w}$ to $1/N$, leading to a good solution even after just a single iteration, as:

$$w_j = \frac{1}{k} \sum_{p=1}^{k} \frac{K(\mathbf{d}_p, \mathbf{C}_j)}{\sum_{l=1}^{N} K(\mathbf{d}_p, \mathbf{C}_l)}. \tag{5.5}$$

Given a codebook, Eqn. (5.5) is thus equivalent to the familiar soft-assignment representation proposed by [167].

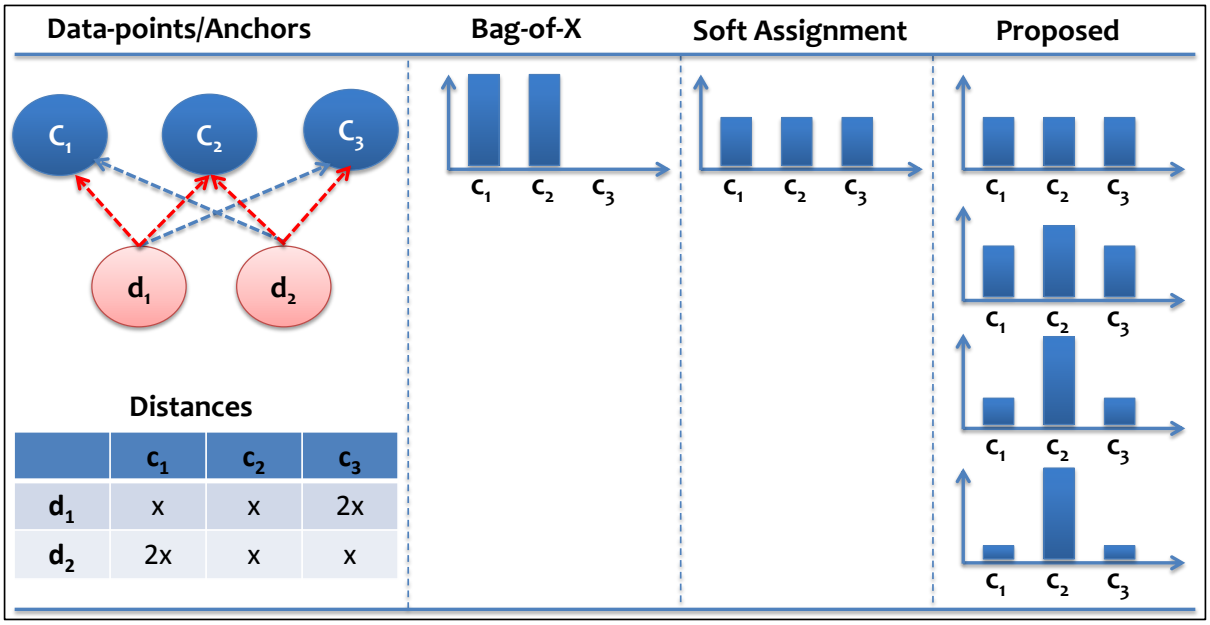


Figure 5.2: A toy example contrasting the proposed representation against traditional BoW and soft-assignment BoW (Codebook Uncertainty [167]). Note that the proposed representation is initially identical to soft BoW but diverges since it maximizes an image-level likelihood score.

5.2.1 A Simulated Example

To contrast the proposed representation against traditional BoW and soft-assignment variants of BoW (such as Codebook Uncertainty [167]), we present a toy example with an "image" containing two features in a 1-D space, a vocabulary with three anchors and a uniform ball kernel (see Fig. 5.2). Traditional BoW simply increments the two bins corresponding to the closest anchors, failing to express the fact that the image contains two similar features. Soft BoW captures this since bin 2 accumulates weight from both features. The proposed maximum likelihood representation seeks anchor weights that optimize the likelihood at an image level (rather than simply accumulating weights). As a result, bin 2 continues to accumulate a greater fraction of weight, resulting in a stronger peak for the shared feature. Algo. 1 summarizes the entire procedure. In practice, even on real data, the algorithm converges in just 3–5 iterations.

5.2.2 *Choice of Kernel Function for Kernel Density Estimate*

For brevity, let us drop the indices from the data-point $\mathbf{d}_p$ and the anchor $\mathbf{C}_j$, to understand the kernel function (K) in detail. A natural choice for a kernel $K(.,.)$ is the Gaussian:

$$K(\mathbf{d}, \mathbf{C}) = \frac{1}{\sqrt{2\pi}r} e^{-\frac{||\mathbf{d}-\mathbf{C}||^2}{r^2}}. \tag{5.6}$$

However, such soft-assignment representations can be unwieldy for large image and video collections because the unbounded support of the Gaussian kernel implies that each visual feature in the image affects the weight corresponding to every anchor. For this computational reason, we advocate the use of bounded support kernels such as a truncated Gaussian or even the simple hyper-ball kernel, which corresponds to a uniform probability of observing a feature in a fixed radius neighborhood of an anchor:

$$K(\mathbf{d}, \mathbf{C}) = \begin{cases} 1 & \text{if } |\mathbf{d} - \mathbf{C}| \leq r, \\ 0 & \text{otherwise.} \end{cases} \tag{5.7}$$

Such a kernel function can be efficiently computed on a large set of anchors, particularly when paired with an approximate nearest neighbor algorithm [8, 110].

Fig. 5.3 illustrates the factors that affect the computation of the weights for any given image using Algo. 1. The anchors that are input to the algorithm can either be taken from a standard clustering-based vocabulary or selected from the ensemble of visual words using random sampling, with a uniqueness constraint to ensure better initialization.

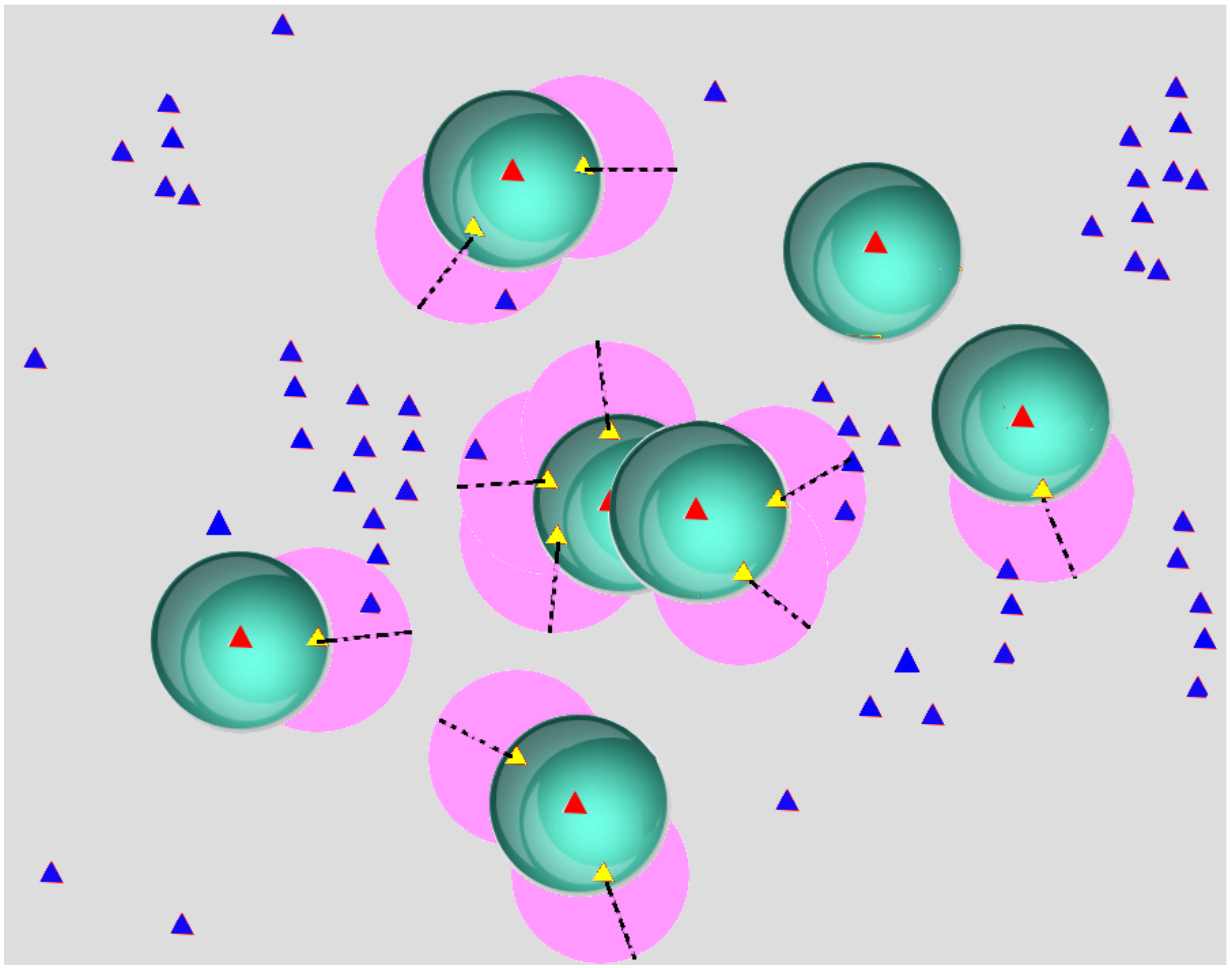


Figure 5.3: A schematic diagram of the proposed procedure. Blue triangular markers indicate all data points (S). Yellow markers denote the *anchors* (C). Purple circles centered at the anchors signify the m-sphere (β_j) that is constructed using the simple kernel function in Eqn. (5.7). Datapoints within these spheres, which have anchors in their r-neighborhood (represented by a green sphere), are indicated as red markers. These datapoints can be viewed as contributors to the representation of the datapoint $\mathbf{d}_p$ through the denominator of Eqn. (5.5).

Procedure ComputeWeights (C, d_p, r)

Input: Set of N anchors (C), Set of M Interest Points (d_p) from p-th instance I_p, Radius of influence (r)

Output: Set of weights $\mathbf{w}$

$\mathbf{w}' \leftarrow \mathbf{0}$;
while *not converged* **do**
  for $j = 1 \ldots N$ **do**
    $n \leftarrow 0$ $i \leftarrow 1$ $w[j] \leftarrow 0$;
    for *each* $d_p[i] \in \{||C[j] - d_p|| \leq r\}$ **do**
      $S_l \leftarrow 0$ $l \leftarrow 0$;
      for *each* $C[l] \in \{||d_p[l] - C|| \leq r\}$ **do**
        $S_l = S_l + w'[l]K(d_p[i], C[l])$;
      $n \leftarrow n + 1$;
      $w[j] \leftarrow w[j] + \frac{w'[j]K(d_p[i],C[j])}{S_l}$;
    Normalize $(\mathbf{w})$;
    $\mathbf{w}' \leftarrow \mathbf{w}$

Algorithm 1: Algorithm to compute $\mathbf{w}$ for a set of data-points extracted from a single image or a video.

5.3 Experiments

We conducted several independent sets of experiments on a standard scene dataset and two widely popular video datasets, namely Scene-15 dataset [83], KTH Human Actions [144] and UCF [100] action datasets. In addition, we used an aerial video dataset that has recently been released by the DARPA VIRAT program. For all these datasets, anchors are generated by sampling $1.2\%, 2.5\%, 5\%, 10\%, 20\%, 40\%$, and 80% of the total number of features in their individual feature ensembles (S).

These anchors are input to our maximum likelihood representation, which depends upon the range search technique we employ in Algo. 1, in particular the search radius which corresponds to the radius (r) of m-spheres $(\{\beta_j\}_{j=1}^{N})$. We use a composite tree indexing scheme (combination of kd-tree and hierarchical k-means) with different search radii to perform the range search required to identify neighbors within the radius of influence of each m-sphere in question. The initial search

radius is obtained by computing the average Euclidean distance between a randomly sampled tenth from S. This measure is further refined by increasing it until a situation is reached where all anchors have quantized at least one feature from S.

Classification is performed using a multi-class SVM with a histogram intersection kernel. This kernel has been shown to perform well in conjunction with bag-of-visual words representations on a variety of datasets, including Scene-15 and Youtube in [83, 100, 167].

5.3.1 Scene-15 Dataset

This dataset consists of a collection of 4,485 images spanning 15 categories, including both natural and man-made scenes. We closely follow Lazebnik *et al.*'s experimental methodology, where we select 100 random images of each category for training and employ the remaining 2,985 images for testing. For all scenes, visual words are extracted using three popular approaches, (a) SIFT [101] on grayscale images, (b) Color SIFT [166], and (c) Gray-SIFT Spatial Pyramid Features [83]. As in Lazebnik *et al.*, we densely sample these descriptors over the image with an 8-pixel stride rather than using an interest-point detector. We use the first two levels of a pyramid with codebook size of 400 while extracting the features, as this was reported to work best.

Fig. 5.5(a) shows a performance comparison of these three features for different sets of anchors. Consistent with earlier work, visual words based on spatial pyramid features perform better than gray SIFT or color SIFT features alone. Our method achieves $75.5 \pm 0.63\%$ accuracy with only 20% of the total number of visual words. We directly compare the proposed representation with our implementations of: (a) standard codebook model with hard clustering, (b) a soft-assignment model (Codeword Uncertainty [167]) using densely-sampled gray SIFT features. In this setting, the anchors input to Algorithm 1 are replaced by cluster centers returned by k-means clustering algorithm. Our results are shown in Fig. 5.5(b). For codebook sizes greater than 1600, our method performs better than the hard and soft assigned codebook models. The main computationally intensive step in our method involves determining the memberships of each anchor, which

we perform using FLANN [110]. The computation of weights using Algorithm 1 is very efficient. A MATLAB implementation of the alogrithm takes less than 5 secs on a standard laptop.

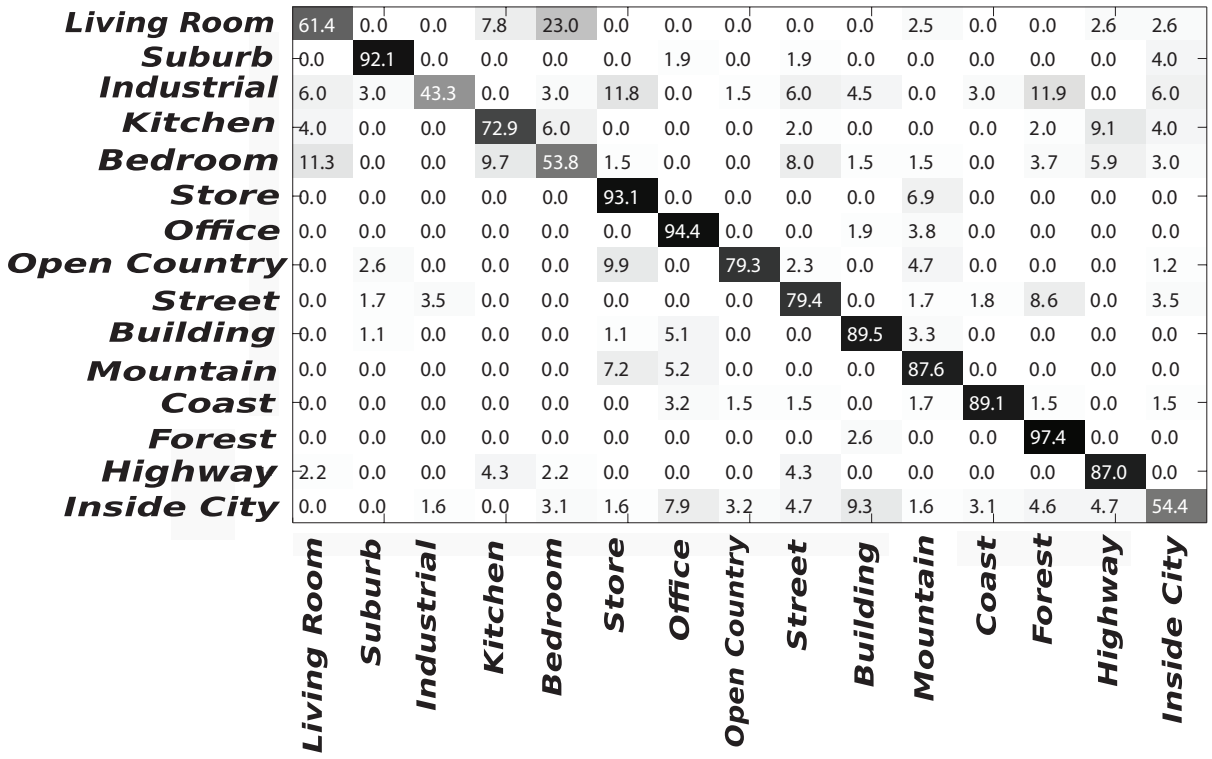
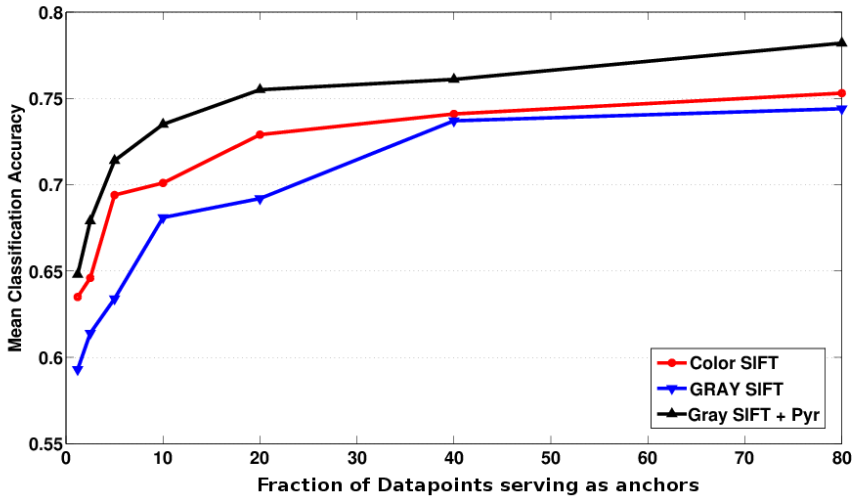
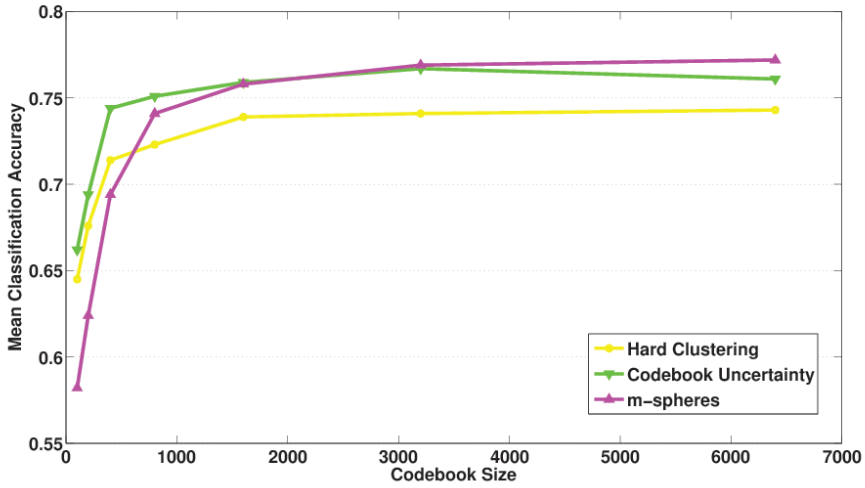


Figure 5.4: **Confusion matrix for the Scene-15 dataset:** The results shown here are based on Gray-SIFT Spatial Pyramid features. These results (mean accuracy per class: $78.8 \pm 0.45\%$) correspond to the maximum likelihood representation generated by using 80% of visual words as anchors. With only 20% of anchors we achieve a mean accuracy of $75.5 \pm 0.63\%$ per class.



(a)



(b)

Figure 5.5: Quantitative analysis of performance of our method in the Scene-15 dataset: 5.5(a) across different feature modalities, namely gray SIFT, color SIFT and spatial pyramid on top of Gray SIFT on vocabularies created using 1.2%, 2.5%, 5%, 10%, 20%, 40%,and 80% of the total number of datapoints from the dataset. Spatial Pyramid features outperform both gray SIFT and color SIFT features. 5.5(b) against different vocabulary construction strategies. The sampling of anchors is replaced by k-means clustering. The x-axis indicates the number of clusters chosen starting from 500 to 6500. The yellow curve shows the performance of standard bag-of-visual-words where the representation is a histogram. The green curve corresponds to bag-of-visual-words with soft assignment proposed in [167]. Our method outperforms both methods at codebook sizes greater than 1600.

5.3.2 *KTH Action Dataset*

The KTH action dataset [144] is a human action dataset that remains popular in the computer vision community. KTH consists of six sets of actions performed by 25 different human actors under four different illumination scenarios. We handpick a set consisting of 598 action clips from all scenarios for our experiments.

Our low-level features are identical to those employed by recent action recognition methods. Each video clip is represented using a collection of datapoints that are extracted in the following manner: (1) Spatio-temporal cuboids are extracted around regions where the detector proposed by Dollar *et al.* [39] produces maximal responses, only a maximum of 200 cuboids are retained per video, (2) Each cuboid is represented by using normalized gradients descriptors, (3) PCA is applied to reduce the feature vector dimension to 100. Thus each video is represented in terms of about 200 visual words, each described by a 100-dimensional vector.

For classification, we build a training set from 10 randomly-selected actors, actions performed by the remaining 15 actors are used as test set. This is repeated 5 times using a multi-class SVM. Since the feature vectors are extremely sparse (as seen in Fig. 5.1(c)), the classification is computationally efficient.

The best average classification accuracies for this dataset are achieved with 10,730 anchors, which is 10% of the total number of visual features. Table 5.1 presents the performance reported by our method and some of the popularly-cited methods in action recognition literature. The accuracy scores are directly imported from the respective authors' papers. A direct comparison is unwise since the experimental methodologies are not identical. However, these results do support our claim that the proposed representation can achieve state-of-the-art performance on standard vision datasets without any explicit tuning. A quantitative confusion matrix is presented in Fig. 5.6 showing the average classification accuracies of each action category using the same set of 10,730 anchors.

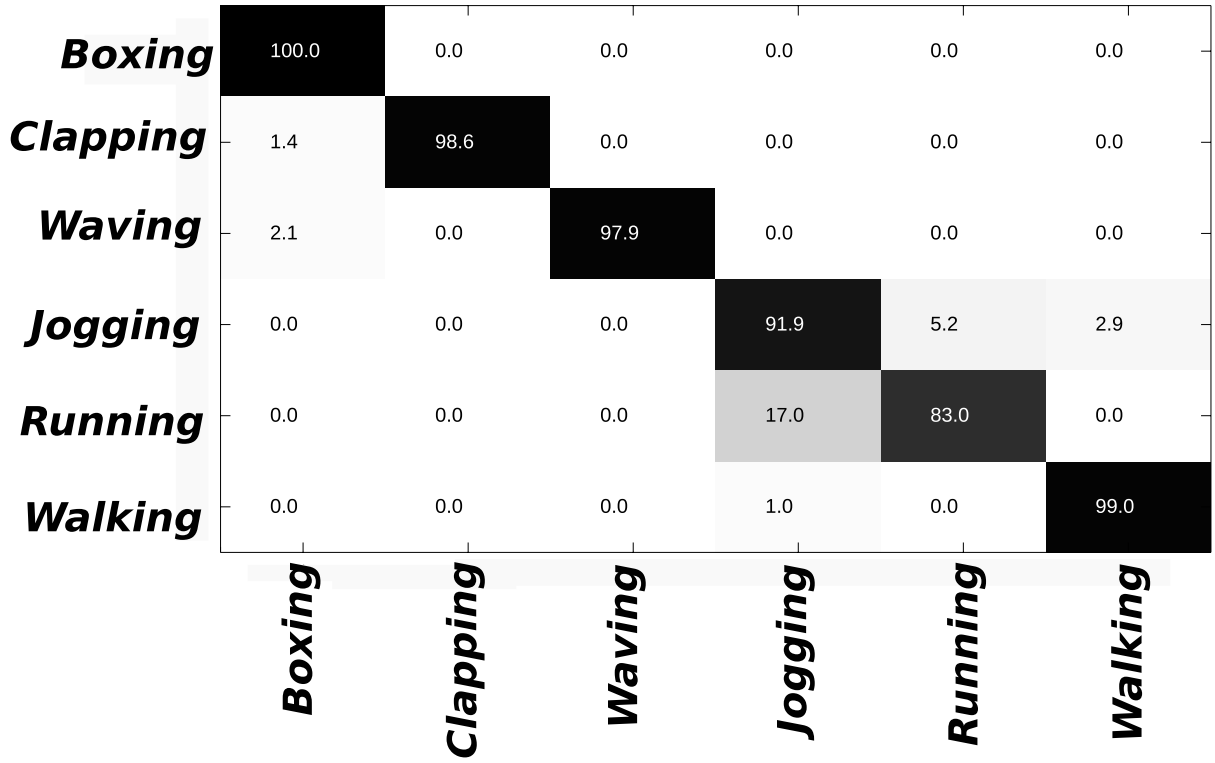


Figure 5.6: Classification results on KTH action dataset with anchors selected from 10% of the total number of video words (avg. accuracy: $95.06 \pm 0.44\%$). The actions Running and Jogging are most confused because of their visual similarities.

Table 5.1: Comparison of the proposed method with published results in action recognition on KTH dataset. We also compare our results with a standard hard-clustering Bag-of-video-words technique that uses k-means clustering to construct its vocabulary followed by SVM for classification.

Method	Mean Accuracy (%)
Proposed method	95.06 ± 0.44
Lin *et al.* [96]	95.77
Liu and Shah [99]	94.15
K-means clustering + SVM	88.34

5.3.3 YouTube Action Dataset

Motivated by the success of our technique on action recognition in KTH, we investigate how our method performs on a newer and more challenging dataset, the YouTube Action Dataset.[1] This dataset is a categorized collection of amateur video clips downloaded from YouTube organized in 11 different categories corresponding to real-world actions such as riding a bicycle/horse/swing, swinging a golf club/tennis racquet, shooting basketball, jumping on a trampoline, juggling a football, diving into a pool, spiking a volleyball and walking with a dog. There are about 100 clips per action. Most of the clips are of poor resolution compared to the KTH data and have noisy and cluttered backgrounds. These clips also exhibit a lot of variation in object scale and viewpoint coupled with significant camera motion. We performed our experiments on the first 10 action instances, distributed over 1051 videos.

Similar to the KTH setup, we extract 400 spatio-temporal volumes from each video, and describe them using gradient features which are further PCA reduced to 200 dimensions. In this case, our ensemble contains a total of 350,693 visual features. Anchors are selected at different granularities $(1.2\%, \ldots, 80\%)$. For each action category, 40 examples are chosen randomly for training, limiting the number of actors to 10. The remaining videos serve as test examples. This process is repeated 10 times. Classification is performed in a similar fashion as covered in the earlier section. As observed in Fig. 5.7, we achieve 76.5% average classification accuracy when 20% of the visual words serve as anchors, beyond which increasing anchors is not beneficial. This shows that the selected anchors are sufficiently representative to express the important aspects of the videos. The confusion matrix (Fig. 5.8) confirms that the proposed approach is effective at classifying actions in unscripted real-world video.

[1] www.cs.ucf.edu/~liujg/action_youtube_naudio.rar

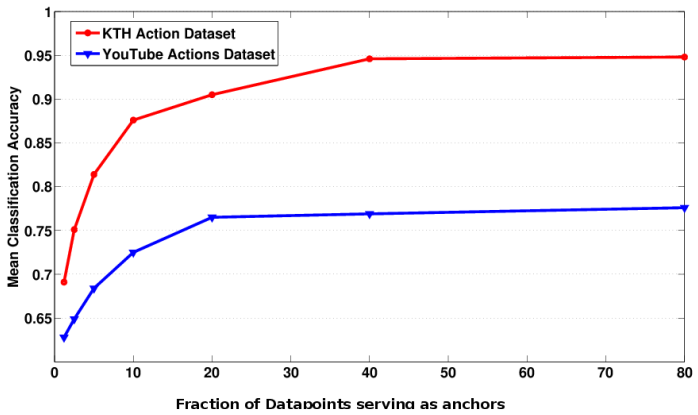


Figure 5.7: Classification results on KTH and YouTube action datasets as the number of anchors is varied. Our accuracy on YouTube ($76.5 \pm 0.8\%$) with ten classes compares favorably with the state-of-the-art results of [100], who report 76.1% on eight classes.

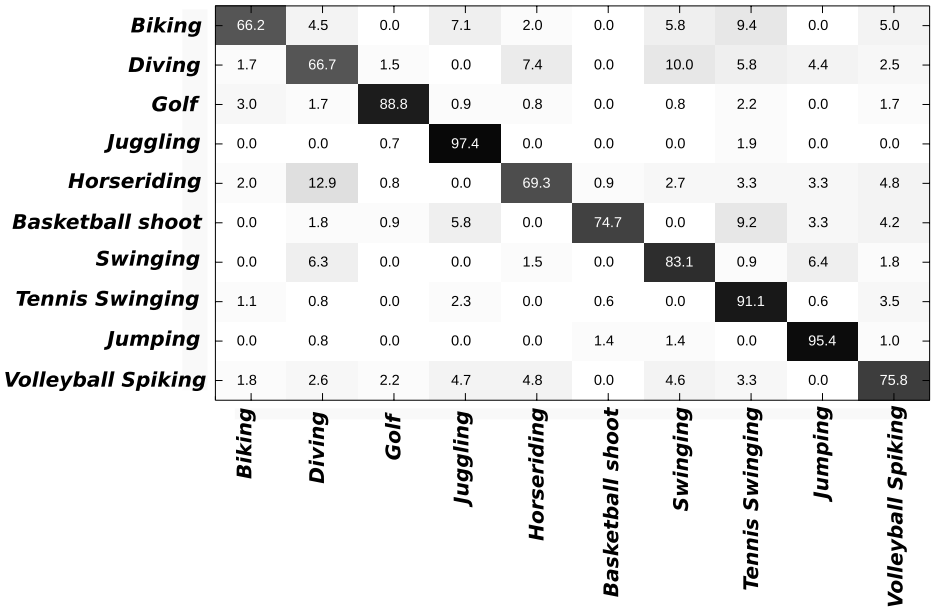


Figure 5.8: Classification results on YouTube action dataset. The mean classification accuracy as determined from the above reaches 76.5%.

5.3.4 VIRAT Aerial Video Dataset

This is a recently released challenging dataset collected under the DARPA VIRAT program consisting of several human and vehicle activities, captured from a moving aerial platform. In this chapter we focus on a subset of the dataset consisting of six human actions. These videos have the following properties that make the action recognition problem in this context more challenging: (1) ego-motion of the camera typically characterized by frequent jitter, (2) extreme low resolution of human actors (50×50 pixels), and (3) large amount of similarity across actions observed from high altitude. For example, the actions standing, gesturing and digging appear similar to each other when viewed from a shaky platform mounted about forty feet above the ground. Similarly, actions such as walking, carrying a box and running can be confused with each other. Each action in the dataset has 200 instances except for gesturing which only has 42 instances.

We extracted two different types of features using two widely popular spatio-temporal feature extraction implementations. In the first setting we used the methodology similar to the previous two experiments on action datasets. In the second, we used Laptev's STIP [81] implementation, which uses a 3-D Harris corner as a space-time interest point detector, with a 144-dimensional concatenation of Histogram of Gradient and Histogram of Optical Flow descriptors. We represent both sets of datapoints using two techniques —the standard bag of video words and the proposed representation. The classification is performed by an SVM with a histogram intersectionkernel using a 10-fold cross validation, similar to the previous experimentalframework. The best performance in this dataset was observed with 388,322 anchors, which is 40% of the datapoints. In this setting, we achieved the maximum mean accuracy of 37.7% per class. Fig. 5.9 shows the confusion matrix. On this challenging dataset, we see from the confusion matrix that the ambulatory actions (walking, running and carrying) can be distinguished from the stationary ones (gesturing, digging and standing). However, there is significant misclassification within these broad categories. We also compare our results with two different types of feature extraction schemes and their respective

bag of words representations in Tab. 5.2. The maximum performance for both the representations are empirically recorded to be at the point where the number of anchors (for the proposed method) or codewords (for BoW) versus mean accuracy becomes asymptotic.

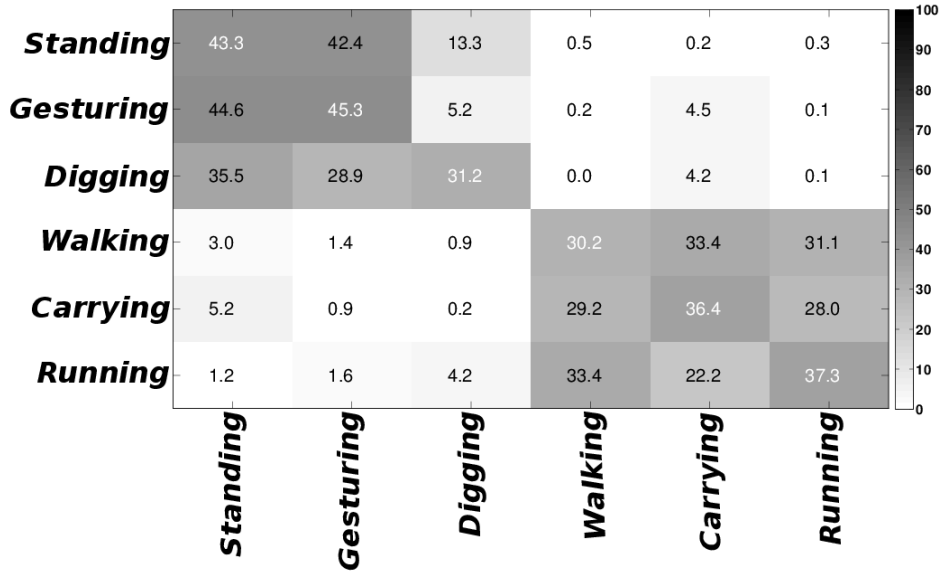


Figure 5.9: Classification results on VIRAT Aerial Video Dataset. Ambulatory actions can be distinguished from stationary actions, but there is still significant confusion within these broad categories.

Table 5.2: Comparative results with two different types of spatio temporal feature extraction/ description techniques, namely HoG-HoF [81] and PCA-G [39] on two representation schemes: standard bag of words and our proposed method.

	BoW		Proposed	
Action	HOGHOF	PCA-G	HOG-HOF	PCA-G
Standing	41.1	39.2	**43.3**	42.2
Gesturing	40.5	41.5	**45.3**	44.9
Digging	**34.9**	34.6	31.2	34.2
Walking	**32.9**	32.6	30.2	31.7
Carrying	35.5	33.7	**36.4**	36.1
Running	34.5	**39.3**	37.4	38.2

5.3.5 Spatio-temporal Concepts Dataset

In order to evaluate the efficacy of our method in detecting a large number of spatio-temporal concepts, the results of which can later be used for complex event recognition, we performed extensive experiments on an in-house dataset constructed from short video clips, approximately $5 - 10$s in duration. These clips are obtained from $2,062$ high definition videos released under TRECVID MED 2011 complex event collection [1]. Details of the events are provided in Chapter 4 and Chapter 6. The following spatio-temporal concepts are annotated from these videos with parentheses including the number of samples obtained for each concept category:

(a) Single person – face visible[84]: Person bending(141), Person blowing candles (94), Person carving(201), Person casting(52), Person cheering(44),Person clapping(147), Person cleaning(22), Person climbing(56), Person closing trunk (14), Person crying (12), Person cutting cake (31), Person cutting fabric(56), Person cutting (195), Person dancing(173), Person dragging(12), Person drilling (97), Person drinking (31), Person eating (155), Person erasing(22), Person falling (98), Person fitting bolt(189), Person flipping (272), Person gluing (14), Person hammering (154), Person hitting (19), Person holding sword (23), Person hugging (115), Person jacking car (46), Person jumping (402), Person kicking (21), Person kissing (143), Person laughing (88), Person lifting (28), Person lighting candle (21), Person lighting (22), Person marching (12), Person measuring (13), Person opening door (21), Person opening trunk (25), Person packaging (21), Person painting (23), Person petting (53), Person picking (21), Person planing(433), Person playing instrument (27), Person pointing (113), Person polishing (39), Person pouring (42), Person pulling out candles (22), Person pushing (58), Person reeling (125), Person riding (11), Person rolling (26), Person running(91), Person sawing (180), Person sewing(177), Person shaking hands (13), Person singing (117), Person sketching (11), Person sliding (189), Person spraying (19), Person squatting (108), Person standing up (19), Person steering (21), Person surfing (62), Person taking pictures (131), Person throwing (74), Person turning wrench (215), Person twist (19), Person twisting wood (16)

, Person using knife (130), Person using tire tube (290), Person walking (421), Person washing (107), Person waving (94), Person welding (23), Person wetting wood (21), Person whistling (12), Person wiping (44), Person writing (98);

(b) Single person – partial body visible[11]: Hands visible(1243), Spreading cream (146), Vehicle moving (547), Wheel rotating (60), Open door (25), Stir (46), Shake (24), Open box (44), Close door(17), Blowdrying(15), Taking pictures (13);

(c) Multiple persons[3] : People marching (337), People dancing (151), People cheering (12);

(d) Non-human[6]: Animal eating (354), Animal approaching (59), Flash photography(17), Machine carving (19), Machine planing (23), Machine sawing(20).

For each video, we investigated three different modalities namely appearance, motion and audio, to extract features to train detectors. In our approach, appearance features are extracted both on a local (SIFT [101]) and a global (GIST [123]) basis. We uniformly sample every 25-th frame from an annotated clip and extracted SIFT (128 Dim.) and GIST (960 Dim.) descriptors from the candidate frames. In order to extract motion information from a video, we employ Dollars [39] cuboid, and Laptev's [81] space-time interest point detectors. The former collects 3-D cuboids around locations where a predefined space-time filter response is significantly high, while the latter uses a 3D Harris corner detector and returns a combination of histograms of oriented gradients and optical flow as a descriptor. Cuboids are later described using optical flow and their dimensionalities are reduced to 250 using PCA. In addition to interest point based methods, we use motion boundary histograms that are trajectory based descriptors and are proven to be more effecttive in action recognition *et al.* [170]. For, audio we extract Mel Frequency Cepstrum coefficients (MFCC) features which represent the short-term power spectrum of an audio signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. MFCC and its first, and second order derivatives are extracted from overlapping segments of 25ms from raw PCM channel obtained after down sampling the source audio signal to 16khz. This results in a $3 \times 21 = 63$ dimensional descriptor for each segment of audio.

All features are independently subjected to quantization in a bag-of-features based representation scheme to produce a final feature for a baseline detector. Finally a one-against-all approach is ensued to train n binary SVM classifiers using a histogram intersection kernel, n being the number of low-level event detectors we intend to train. Thus for a given vocabulary size, we obtain $6 \times n$ different SVM classifiers. This methodology is repeated for the proposed anchors based method [21] yielding $6 \times n$ different SVM classifiers. We quantitatively analyze the performance of our low-level spatio temporal concept detectors using two metrics: (a) Area under Detection Error Trade-off (DET) curves [107], and (b) Average precision. In the former, a given classifier's success is measured in terms of miss-detection and false alarm probabilities, while the latter is more meaningful in terms of true detection and false alarm rates. DET is considered to be more effective when the proportion of negative samples to positive samples in the test data is large and is a standard evaluation metric used by NIST.

Fig. 5.10 visualizes the mechanism of typical low-level event detectors over any arbitrary video. Each video is automatically divided into fixed length clips, on which low-level event detectors are applied. In the above figure, a typical video depicting the event "Changing a tire" is used as an example. The responses these detectors using motion (red), static (green) and audio (blue) modalities on relevant temporal segments of the video are shown qualitatively. A simple weighting scheme is applied on the detector confidences to achieve the final decision on the presence of a concept. In the following sections, we discuss our experiments in detail and show quantitative results.

It is interesting to observe that for all feature modalities, our proposed representation outperforms the popular bag-of-words based representation by $2 - 4\%$. This gain is significant as the number of concept categories detected are sufficiently larger than standard datasets and hence makes a strong case in favor of the proposed representation for large scale visual recognition tasks. A detailed AP measure for 62 different concept detectors using different features, ultimately transformed to the MLE based representation is provided in Fig. 5.11.

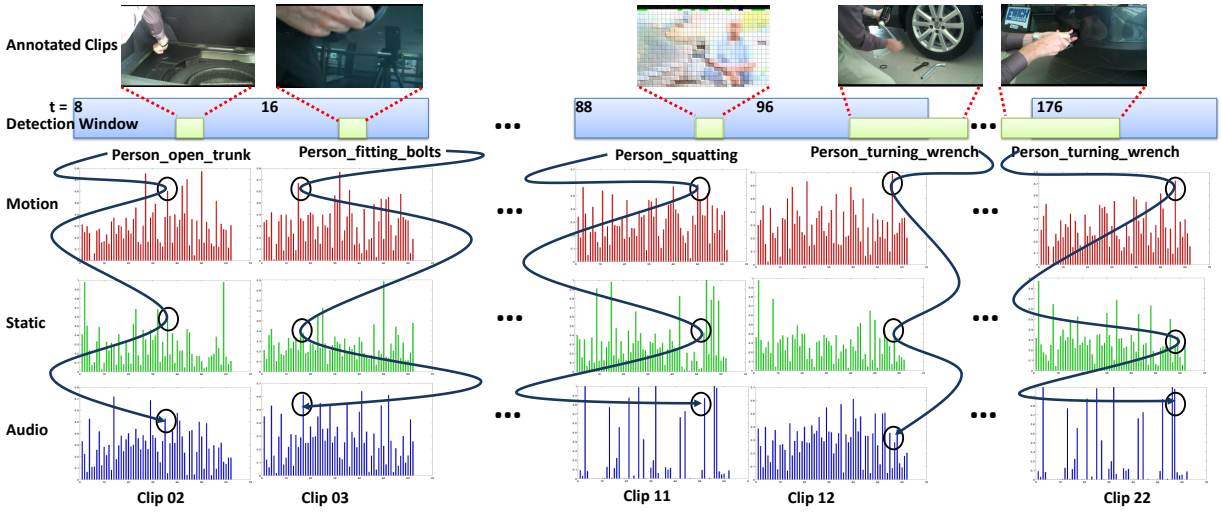


Figure 5.10: **Low-level event detectors "in action":** The figure shows the responses of low-level event detectors in an arbitrary video depicting the complex event: changing a tire. The top row shows sampled frames from a given video. The blue horizontal bar gives a sense of the temporal sampling window (in this case it is 10s with 2 seconds overlap) on which pretrained low-level event detectors are applied. The smaller green bars correspond to the actual granularity of the low-level event (obtained from annotation). The bottom 3 rows show the detector responses from different feature modalities. After combining the responses of all the detectors from different modalities, we observe that the low-level event: "person open trunk" is detected with maximum confidence in the shown window. This is very close to the ground truth. Similar trend is observed for different low-level events like : "person fitting bolts", "person squatting" and "person turning wrench", which are all very relevant to the event "changing a tire".

5.4 Summary

We presented a novel, principled representation [21] for both images and videos that is based on maximizing the likelihood of generating the observed visual words using a vocabulary. We introduced a computationally efficient iterative algorithm that identifies the globally optimal parameters. Recent approaches that employ soft assignments are shown to be special cases of our approach, and our method is completely compatible with recognition systems that operate with standard bags-of-visual words representations. Furthermore, we show how the expensive step of clustering visual words to generate a vocabulary can be replaced (for our representation) with a sampling-based approach over visual words without significantly impacting classification accuracy, and in some case performing better than bag-of-features based approaches.

Table 5.3: **Spatio temporal concept detector performance evaluation summary:** This table summarizes the performance of detectors constructed from BoW (Baseline) and our proposed anchors based method, across different feature modalities using two different metrics – average area under DET curve (Avg. AUC), and average precision (AP). Both the metrics provide a coarse idea of the performance of the detectors trained on 62 spatio-temporal concepts. The lower the AUC measure (on a scale of $0-1$) the more reliable the detector is. This is in contrast to the AP measure, for which the greater the score the better the performance. A more detailed comparison is provided in Fig. 5.11.

	Representation			
	Bag-of-Features		**Anchors**	
Modality	**Avg. PAUC**	**AP(%)**	**Avg. PAUC**	**AP(%)**
Static [SIFT]	0.2203	22.01	0.1949	23.52
Static [GIST]	0.2718	18.21	0.2223	18.15
Motion [Dollar]	0.1948	16.22	0.1735	19.24
Motion [STIP]	0.1869	17.31	0.1878	19.42
Motion [MBH]	0.1721	19.21	0.1639	20.21
Audio [MFCC]	0.3121	11.13	0.2936	11.12
Fusion [SIFT+MBH]	0.1615	19.41	0.1524	21.02

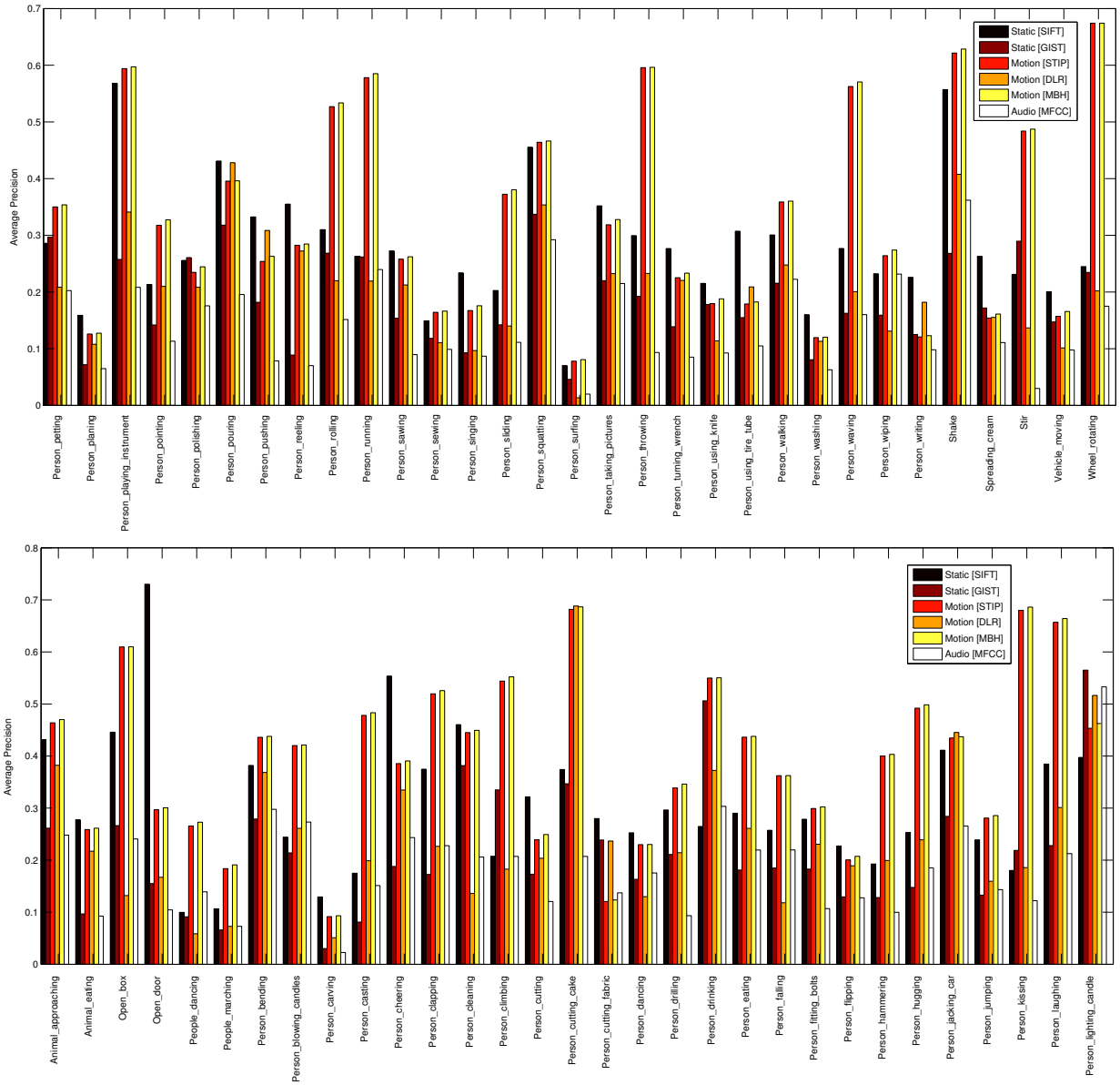


Figure 5.11: **Spatio temporal concept detector evaluation:** The average precision for different concept detectors using our anchors based representation on top of appearance (SIFT, GIST), motion (STIP, Dollar, MBH) and audio modalities (MFCC) are shown.

CHAPTER 6: UNDERSTANDING TEMPORAL DYNAMICS OF ACTION CONCEPTS FOR COMPLEX EVENT RECOGNITION

6.1 Introduction

Temporal interactions between concepts have been represented using graphical models (directed and undirected) in the past extensively by researchers using Hidden Markov Models [92,184, 187], Bayesian Networks (BNs) [59, 62], Conditional Random Fields (CRF) [35, 165, 172, 176], Dynamic Bayesian Networks (DBNs) [60] etc. While these models are mathematically elegant, most of them need extensive domain specific knowledge in addition to a large number of training samples, apart from being sensitive to underlying representations, which may incorporate noise. In this chapter, we propose an alternative technique to model such temporal interactions between spatio-temporal concepts from the perspective of linear dynamical systems, which generalize some of the popular graphical model based approaches.

Our work emphasizes on extracting joint temporal evolution of underlying models which can be used in the recognition of complex events. In our approach, a video is decomposed into a sequence of fixed-length temporal clips, on which low-level feature detectors are applied. Each clip is then represented as a histogram (bag-of-visual-words) which is used as a clip level feature and tested against a set of pre-trained action concept detectors. Real valued confidence scores, pertaining to the presence of each concept is recorded for each clip, reducing the video into a vector-time series. Two sets of novel features are computed on the vector time-series which capture temporal relationships between different concepts using Linear Dynamical System (LDS) theoretic approaches. The first being principal projections of a block Hankel Matrix constructed directly from the vector time series. The second corresponds to joint characteristics of time-series such as lag-invariance, frequency proximity, periodicities etc. Together they form a global feature (Fig. 6.1), which is plugged into a discriminative classification algorithm for recognition of complex events.

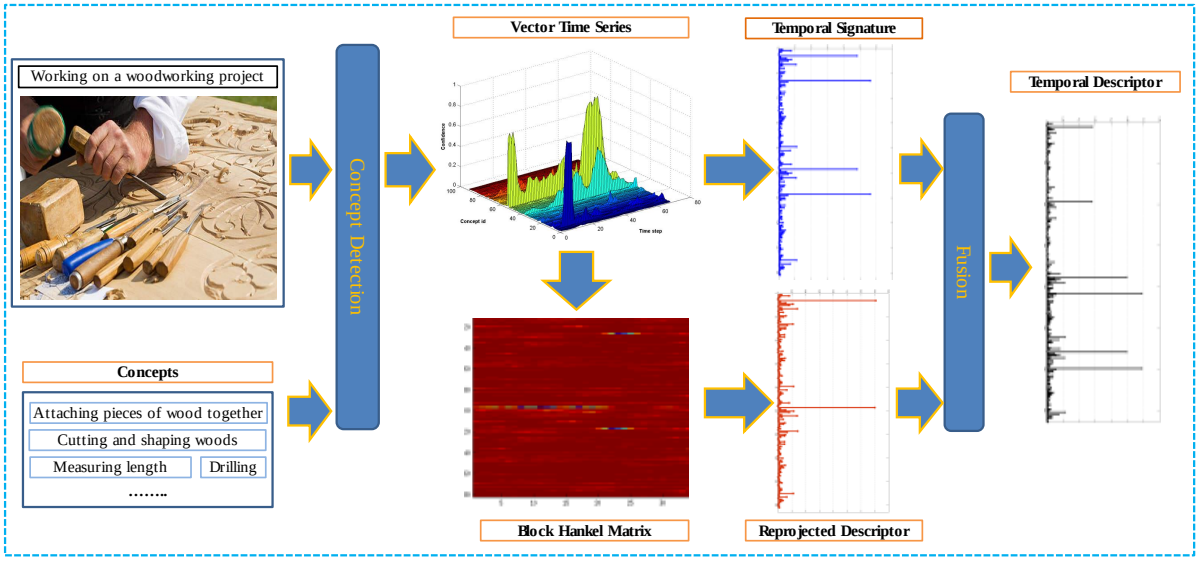


Figure 6.1: **A schematic diagram showing** the various stages involved in our proposed temporal feature extraction from a typical video.We build our complex event recognition computational pipeline (discussed in Sect. 6.3)based on the above methodology. Please refer to text for a detailed explanation.

Although our formulation of a complex event is fundamentally based on LDS represented by first order Markov chains, we circumvent the problems encountered by methods that enact similar principles (HMMs and their variants) by computing features from LDS that do not perform direct estimation of the parameters (priors, observation matrices, and, transition matrices). Specifically, we make the following contributions in this chapter: (1) We introduce two different sets of algorithms to the multimedia community, that can be used to extract features from any vector time-seres data, (2) We demonstrate the superiority of our algorithms over others [92, 184, 187] that have been predominantly used in recognizing sequential multimedia data through exhaustive experimental evaluation, (3) We compare our proposed technique with current state of the art techniques that do not incorporate temporal information, and show respectable improvements.

Our proposed method is also conceptually more intuitive than earlier approaches that rely on intermediate concept-based representation schemes [67, 113] to model complex events. As

explained in Fig. 6.2, our intermediate representation is constructed from joint statistics of individual concepts and hence is more robust in dealing with errors introduced by the individual detectors. Our experiments provide conclusive evidence in favor of extracting joint evolutionary statistics over simple statistical max-pooling or average-pooling over sequential data. In addition to the computational benefit we achieve in extracting these features, our time-series based feature extraction scheme requires very limited domain knowledge – to the extent of the intermediate concept representation level, thereby reducing the input parameter space (in terms of explicit domain knowledge) as required by DBN [60] and CRF based methods [35, 172]. Finally, our technique is flexible enough to be integrated into any intermediate representation as long as it is sequential.

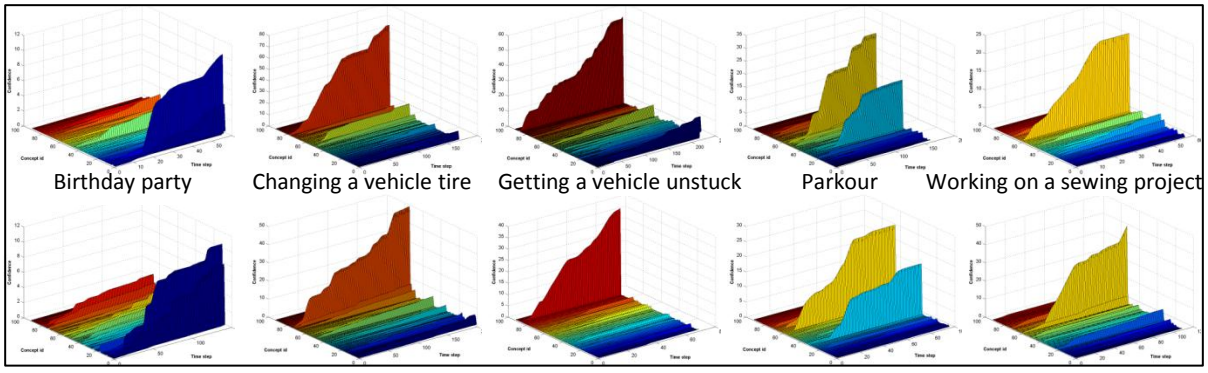
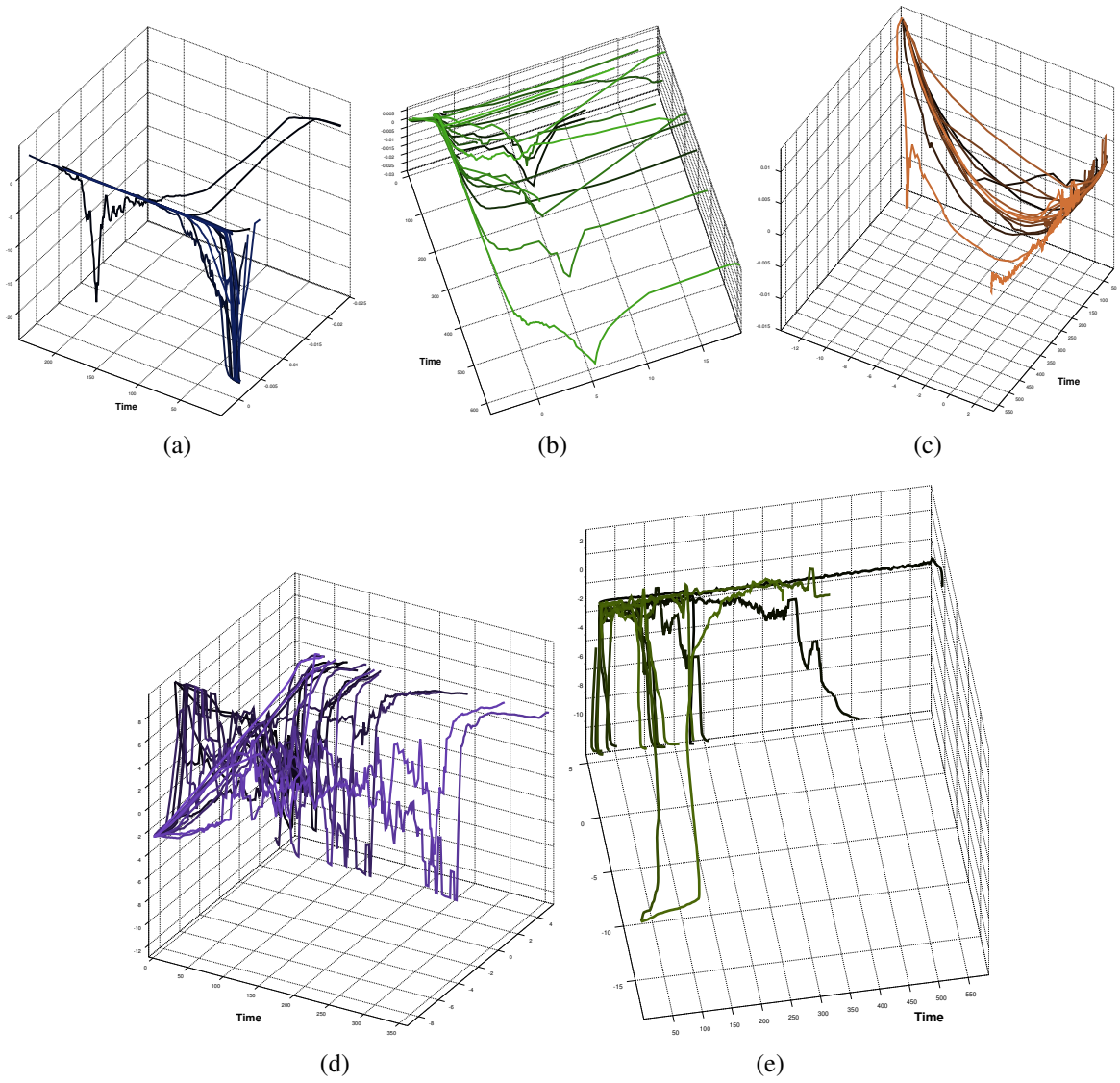


Figure 6.2: **Temporal evolution of concepts in complex events:** Cumulative density distributions of concept detector confidences for 2 different video samples (rows) from 5 event categories (columns) in TRECVID MED 2011 dataset. Each color corresponds to a different concept, while the evolution of each concept is shown as a cumulative distribution of its confidence (z-axis) over different time-steps. This figure shows how a concept slowly evolves along the duration of a video and how evolutions show some discriminative pattern across videos of same event categories. More samples are shown in an alternative representation in Fig. 6.3.

For the sake of legibility, we organize this chapter into the following sections: In Sect. 6.2, we provide an overview of our approach, discussing the computation of Hankel matrix based features in Sect. 6.2.1 and temporal signatures in Sect. 6.2.2, respectively. This is followed by Sect. 6.3, where we provide the experimental details. Next, in Sect. 6.4, we produce our results which is followed by a detailed discussion in Sect. 6.5. Finally, we summarize this chapter in Sect.

6.6 with some insights towards future work.



(a) (b) (c)

(d) (e)

Figure 6.3: **LDS representation for video samples pertaining to** 5 **event categories**: Each trajectory in a figure depicts an LDS generated for a particular event from a videos. Each figure has 80 trajectories color-coded to show separation. Each video, originally a p-dimensional vector-time series of p concept detector responses, is reduced to a sequence of 2-dimensional points for visualization purpose with (a) showing the event "Birthday party" (E006), (b) "Changing a vehicle tire" (E007), (c) "Getting a vehicle unstuck" (E009), (d) Parkour (E013), and finally,(e) Working on a sewing project (E015). It is to be noted that each trajectory is mostly smooth even with the dimensionally reduced representation which argues in favor of our choice of LDS for modeling them. However, the shapes of the trajectories are not discriminative for an event class, so clustering them in this space is not meaningful. This motivates us to chose our temporal features (shown in Fig. 6.4, 6.5).

6.2 Approach

In our formulation, we represent a video V as a sequence of fixed-length clips $S_0, S_1, \ldots, S_{n-1}$, with a certain number of overlapping frames, n being the total number of clips in the given video. On each clip, a fixed set of concept detectors are applied and their respective responses denoting the probability of presence of the corresponding concepts, are recorded. Thus $V \equiv \{\mathbf{c_0}, \mathbf{c_1}, \ldots, \mathbf{c_{n-1}}\}$, where each $\mathbf{c_t} \in \mathbb{R}^p$, is a vector containing concept detector responses p being the number of concepts. Therefore, each corresponding observation can be independently considered as a temporal sequence $\forall \mathbf{c_t} \in V$, depicting the evolution of a particular concept throughout the video. A natural choice is to treat each of these time-series independently, for either fitting in an appropriate statistical model [175] or extracting statistical features. However, these techniques are too simple to capture the interactions between individual time-series of concepts.

Assuming each vector of concept detector responses is directly observed to be emanating from a slowly evolving process (refer to Fig. 6.2 and Fig. 6.3 for ease of understanding), using foundations from Linear Dynamical Systems, we can describe it with the following set of equations:

$$\mathbf{c_t} = K\mathbf{x_t} + \epsilon_\mathbf{t}, \tag{6.1}$$

$$\mathbf{x_t} = \phi\mathbf{x_{t-1}}; \quad \mathbf{x_0} \quad \text{given}, \tag{6.2}$$

where, K is the observation matrix $\in \mathbb{R}^{p\times\theta}$ that maps each observed time-step to a relatively lower dimensional hidden state vector $\mathbf{x_t} \in \mathbb{R}^\theta$, $\epsilon_\mathbf{t} \sim \mathcal{N}(0, 1)$ (noise), and ϕ is the dynamics or transition matrix $\in \mathbb{R}^{\theta\times\theta}$ which relates the current hidden state with the previous hidden state. Analogously, the dynamics matrix ϕ predicts the hidden states for the next time-step, while the observation matrix K provides us with the information on how the hidden variables are mapped to the observed concept detector responses at each time step, with each row of K corresponding to a

concept sequence.

Thus, a system defined using Eqn. (6.2) can be identified efficiently if one can directly estimate the parameters K, ϕ and $\mathbf{x_0}$. However, this is an ill-posed problem [88] as K and ϕ are not unique for a given set of sequences, and are usually subject to permutation, rotation and linear combinations. Thus each row in K cannot uniquely identify the characteristics of the corresponding concept sequence. We investigate two independent techniques, discussed next, which can be used to obtain a discriminative feature for our dynamical system in Eqn. (6.2) without directly estimating K, ϕ.

6.2.1 Block Hankel Matrix Descriptors

Subspace state space system identification techniques [124] attempt to estimate the parameters K, ϕ and $\mathbf{x_0}$ by determining the hidden state sequences through the projection of input and output data. In [124], the authors propose an algorithm to identify the subspace of the hidden states using Hankel matrices, which we extend in the following way to obtain compact features characterizing video specific dynamical systems.

Given a vector time-series V, we can construct the corresponding block Hankel matrix H, as follows:

$$H = \begin{pmatrix} \mathbf{c_0} & \mathbf{c_1} & \mathbf{c_2} & \dots & \mathbf{c_{n-r}} \\ \mathbf{c_1} & \mathbf{c_2} & \mathbf{c_3} & \dots & \mathbf{c_{n-r+1}} \\ \mathbf{c_2} & \mathbf{c_3} & \mathbf{c_4} & \dots & \mathbf{c_{n-r+2}} \\ \dots & \dots & \dots & \dots & \dots \\ \mathbf{c_{r-1}} & \mathbf{c_r} & \mathbf{c_{r+1}} & \dots & \mathbf{c_{n-1}} \end{pmatrix}, \tag{6.3}$$

where r is an integral estimate on the number of entries of the j-th column vector that are sufficient to express the subsequent $(j+1)$-th column of H. Under the same assumptions that our observation vectors are measured from a slowly evolving process (refer to Fig. 6.2), the matrix in Eqn. (6.3) is of rank $s \leq r, n-1-r$. This hypothesis enables us to describe any given observation vector $\mathbf{c_t}$ in

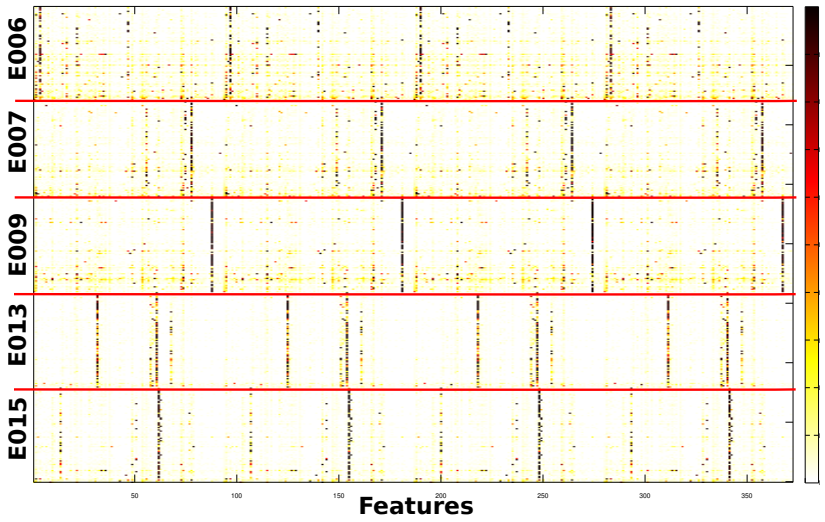
Eqn. (6.2) as an autoregressive model of order s such as:

$$\mathbf{c_t} = k_1\mathbf{c_{t-1}} + k_2\mathbf{c_{t-2}} + \ldots + k_s\mathbf{c_{t-s}}. \tag{6.4}$$

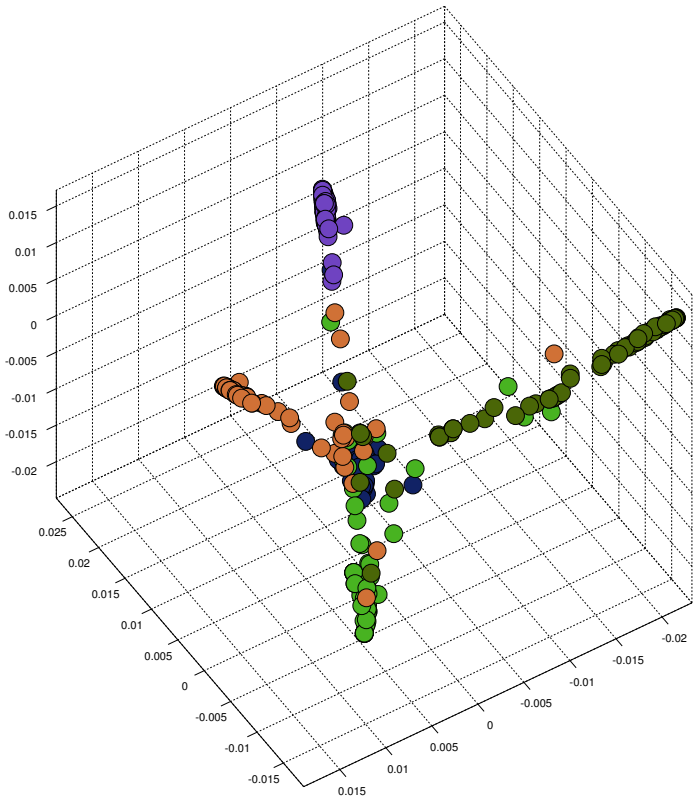
The set of coefficients from all auto-regressors of the form in Eqn. (6.4) conforming to our LDS in Eqn. (6.2) can be obtained by taking projections of HH^T on the m largest eigenvectors after performing singular value decomposition (SVD) on HH^T. In practice, all Hankel matrices are normalized using the Frobenius norm i.e.

$$\hat{H} = H/\operatorname{Tr}(HH^T)^{\frac{1}{2}}, \tag{6.5}$$

in order to make the estimation more robust to noise. The projections are concatenated to create the block Hankel matrix based descriptor from the vector time series.



(a)



(b)

Figure 6.4: **Discriminative representation of LDS in Hankel Matrix feature space:** Hankel matrix based features of samples from the same 5 event categories in Fig. 6.3 are shown in two alternative visualizations. Each row in (a) is feature extracted from a video, grouped into 5 event categories using overlaid dashed yellow lines, while each column is a feature. Note some columns are discriminative to particular events. (b) shows features corresponding to each sample as a circle, similar in color to trajectories shown in 6.3. Note, even in this low dimensional visualization how well samples belonging to same event categories naturally cluster together .

The obtained descriptor captures the interaction between concept sequences nicely, which enables us to obtain a discriminative representation in the proposed feature space. This is shown in Figs. 6.4(a), 6.4(b). Note how some columns similar values within the same event class, and videos belonging to same events cluster naturally. With that said, it is often useful to obtain features that quantify meaningful relationships between temporal sequences, e.g. lag-independence (two time-shifted sequences should be grouped together), frequency proximity (sequences with similar frequencies should be grouped together), harmonic clusters (sequences with similar periodicities should be grouped together). In order to obtain such meaningful relationships between vector-time series, we employ a similar technique proposed in [89], which the authors use to cluster data from one dimensional, fixed length temporal sequences. Our motivation to use [89] for modeling the interactions between evolving spatio-temporal concepts is strongly because: (1) it is proven to have more discriminative power over DCT coefficients obtained from wavelet analysis, and (2) is computationally more efficient than dynamic time warping based methods which have quadratic complexity on sequence length. We hereafter refer to these features as temporal signatures and discuss how they are extracted, in the next section.

6.2.2 Temporal Signatures

Recall Eqn. (6.2) representing the dynamics of a vector time-series. The hidden state vector, $\mathbf{x_t}$ can have limited degrees of freedom depending on the nature of the eigenvalues ($\{\lambda\}_{i=1}^{L}$) of the dynamics matrix ϕ, e.g. exponential growth ($|\lambda_i| > 1, \forall \lambda_i \in \mathbb{R}$) or decay ($|\lambda_i| < 1, \lambda_i \in \mathbb{R}$), stationary sinusoidal periodicity ($\lambda_i \in \mathbb{C}$), and mixtures of all [63]. Since, these eigenvalues encapsulate the signal structure (frequencies, amplitudes, phase) of $\mathbf{x_t}$, they can be obtained by SVD (valid $\forall \lambda_i \in \mathbb{C}$). Thus without any loss of generality, we can write:

$$\phi = U \Lambda U^T, \tag{6.6}$$

where Λ is the diagonal matrix of eigenvalues, grouped into their conjugate pairs and ordered according to their phases, U contains the corresponding eigenvectors. Thus the SVD of ϕ enables us to canonicalize the hidden state variables as:

$$\hat{\mathbf{x}}_{\mathbf{0}} = U^T \mathbf{x}_{\mathbf{0}}, \tag{6.7}$$

$$\hat{\mathbf{x}}_{\mathbf{t-1}} = U^T \mathbf{x}_{\mathbf{t-1}}; \tag{6.8}$$

and compensate the observation matrix (K) to obtain observation vectors pertaining to the dynamics matrix as : $K_h = KU$, where, K_h is the harmonics mixing matrix. Thus, using the above relation in Eqn. (6.8) the hidden state variables and the observation vector can be expressed as:

$$\hat{\mathbf{x}}_{\mathbf{t}} = \Lambda^{t-1} \hat{\mathbf{x}}_{\mathbf{0}}, \tag{6.9}$$

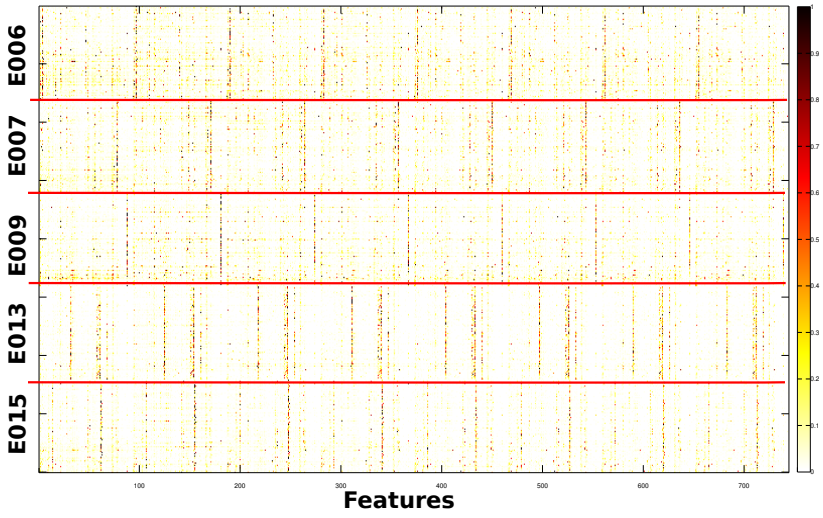
$$\mathbf{c}_{\mathbf{t}} = K_h \Lambda^{t-1} \hat{\mathbf{x}}_{\mathbf{0}} + \epsilon_t. \tag{6.10}$$

Similar to [89], we use an Expectation-Maximization based algorithm initialized with Eqn. (6.10) to determine the final harmonic mixing matrix (K_h) for each vector time series. After a finite number of iterations, we obtain $\hat{K}_h$ which in practice, is of lower rank than K_h, the rank directly corresponds to the number of data-dependent frequency groups conforming to either of growing, decaying or stationary sinusoids.

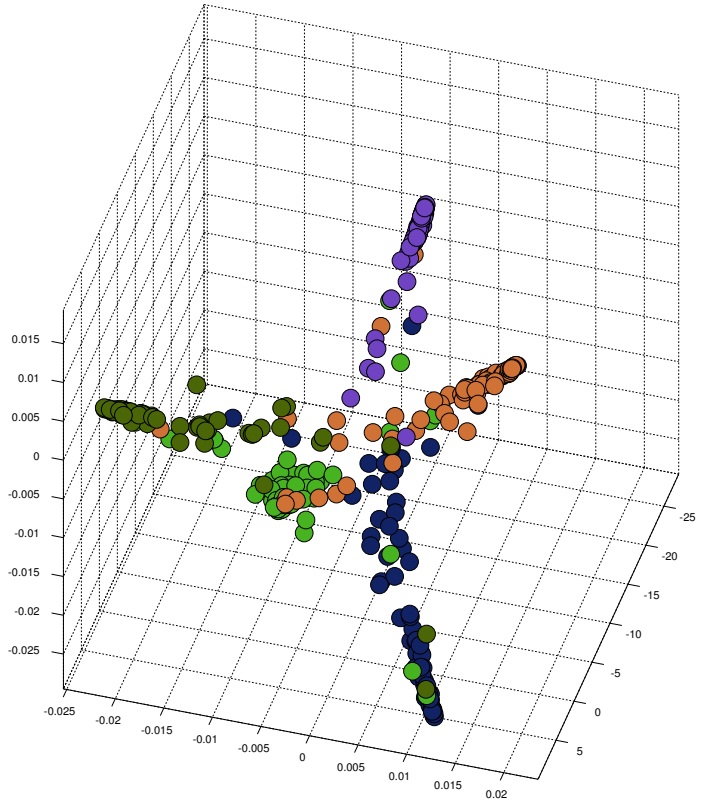
Since entries in the harmonic mixing matrix ($\hat{K}_h$) are complex, the magnitude of this matrix, K_m yields features that are lag independent. These features can be easily computed by performing SVD on K_m yielding matrix of eigenvectors (U_m) and diagonal matrix (D_m). Finally features can be computed by concatenating the rows of the projection matrix ($P = U_m D_m$), each row of which can be interpreted as encodings of frequencies discovered for a particular concept detector response sequence.

The matrices shown in Figs. 6.4(a) and 6.5(a) show how effectively only a few columns

capture discriminative information specific to different event classes. This is also reflected in Figs. 6.4(b), and 6.5(b) which show how even a naive dimensionality reduction technique when applied to the features from LDS, can lead to near-optimal clustering.



(a)



(b)

Figure 6.5: **Discriminative representation of LDS in Temporal Signature feature space:** Temporal signatures of samples from the same 5 event categories in Fig. 6.3 are shown in two alternative visualizations. Please refer to Fig. 6.4 for detailed interpretation..

6.3 Experiments

In order to obtain a meaningful intermediate representation of a video, we first decompose it into a sequence of clips containing spatio-temporal concepts. As discussed in Sect. 6.1, these concepts can be interpreted as unit actions that may or may not be repetitive, typically spanning across $100 - 200$ frames. A set of 93 unique concepts are identified by parsing the textual definition of events provided within NIST's TRECVID MED 2011 database (available to MED11 participants). As an example, the following concepts: *Person clapping*, *Person blowing candles*, etc. are identified for the event *Birthday Party*. A complete list of concepts is attached in the supplementary material.

Using human annotators, we obtain approximately equal number of training examples which are clipped from a portion of original TRECVID MED 2011 dataset, per concept category. Dense trajectory based spatio-temporal features, which report state-of-the-art in action recognition [170] are extracted from each annotated clip, which are further reduced to a bag-of-visual-words (BoVW) representation. A vocabulary size of 2048 is observed to deliver best performance, and hence chosen as default vocabulary for successive experiments. Binary SVM[1] classifiers with histogram intersection kernels are used as our concept detectors. These concept detectors are applied on BoVW representations of each fixed-length clips (300 frames with an overlap of 60 frames) from every video. Next, 93 normalized confidences are collected from each clip leading to a vector time series for every video.

6.3.1 Datasets

A number of datasets have been released by NIST as part of TRECVID MED competition organized since 2010 [2]. We have selected two datasets for our evaluation. The first one released in

[1] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/

[2] http://www.nist.gov/itl/iad/mig/med.cfm

2011, here in referred as MED11 Event Collection (MED11EC), consists of 2062 videos from 15 different complex event categories. These are: *(E001) Attempting a board trick*, *(E002) Feeding an animal*, *(E003) Landing a fish*, *(E004) Wedding ceremony*, *(E005) Working on a woodworking project*, *(E006) Birthday party*, *(E007) Changing a vehicle tire*, *(E008) Flash mob gathering*, *(E009) Getting a vehicle unstuck*,*(E010) Grooming an animal*, *(E011) Making a sandwich*, *(E012) Parade*, *(E013) Parkour*, *(E014) Repairing an appliance*, and, *(E015) Working on a sewing project*.

Similar to MED11EC, in 2012, NIST released a second dataset containing $2,000$ videos pertaining to 10 different events. These are listed as follows: *(E021) Attempting a bike trick*, *(E022) Cleaning an appliance*, *(E023) Dog show*, *(E024) Giving directions to a location*, *(E025) Marriage proposal*, *(E026) Renovating a home*, *(E027) Rock climbing*, *(E028) Town hall meeting*, *(E029) Winning a race without a vehicle*,and *(E030) Working on a metal crafts project*.

All videos in the two datasets are approximately uniformly distributed over the 25 event classes, and are typically recorded by amateur consumers approximately at 30 fps with no specific resolution, under unconstrained scenarios. Also videos from these events have large degree of intra-class visual variance (e.g. attempting a board trick refers to both snow boarding and skate boarding), and in many cases demonstrate subtle inter-class visual variance (e.g. cleaning an appliance and repairing an appliance).

We performed our experiments on two broader dataset settings. The first one involved only the videos in MED11EC, while the second one involved videos in both MED11EC and the $2,000$ videos released in 2012. We refer to this combination as MED12EC which consisted of 25 events distributed across $4,062$ videos.

6.3.2 Baseline Methods

Since, the datasets are relatively new and efforts have only began to be made using concepts, it is very difficult to compare our methods with other TRECVID MED 11 submissions [30, 113, 158, 189], that involve fusion of multiple low-level feature representations. Also, to our

knowledge, none of the published methods have studied temporal interactions between intermediate feature representations. In all our experiments, we report average precision (AP) which is widely accepted in similar recognition tasks [30, 189]. We chose binary linear SVM classifiers to report the performance on event recognition. In all cases, we perform 5-fold cross validation while reporting the average precision.

In order to compare our proposed temporal representation with methods that efficiently model temporal interaction, we implemented 3 independent baseline methods which have been used extensively in event recognition literature [92, 187]. These methods can be used model temporal interactions between spatio-temporal mid-level representations – specifically, vector time series of spatio-temporal concepts. The first one involves computation of a number of Discrete Cosine Transform (DCT) coefficients on each concept detector response sequence. DCT coefficients capture useful frequencies from a waveform (which in our case, is a single concept detector response sequence) in Fourier domain, and coefficients from all sequences from a video can be concatenated to form a vector which forms the final temporal descriptor for classification using a linear SVM. We experimented with 5 sets of DCT coefficients ranging from $8, 16, \ldots, 128$. The best performance was achieved for 64 coefficients per time-series, which required a 93×64 dimensional feature per video. A detailed comparative analysis is provided in Tab. 6.1.

In the next settings, we experiment with two different recognition strategies using first order hidden Markov models which reflect our LDS formulation in section 6.2. The first strategy includes a discrete HMM as a classifier, while the second one includes a continuous HMM [3]. In both cases we perform experiments with 6 different number of hidden state variables (θ) from $2, 4, 8, 16, 32, 64$. Initial parameters for both experimental settings (refer Eqn. (6.2)), the prior ($\mathbf{x_0}$), transmission (ϕ) and observation probabilities (K) are determined from a stochastic process input with $\mathcal{N}(0, 1)$. For discrete HMMs, we obtain the maximum confidence at each time step from every observation ($\mathbf{c_t}$) in a given vector time series, and associate the corresponding concept label to

[3] http://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html

generate the input symbol sequence. This discretization step is not required in a continuous HMM framework, as each state variable in this case is modeled using the distribution of confidences at each time step. In both cases, event specific models are generated, and given a testing sequence, the maximum likelihood of generating the input sequence given an event model is computed using a forward Viterbi algorithm, which identifies the true class of the given sequence.

Fig. 6.6 draws a comparison of our proposed temporal signatures extracted from the $93-D$ vector time-series representations of videos in MED11EC dataset, against two other baseline methods that also involve formulation of LDS, similar to our method. The mean Average Precision scores are shown against number of hidden states ($\theta = 2, 4, 8, 16, 32, 64$) for learning LDS parameters using three similar approaches, with green – Temporal Signatures (TS), purple – discrete HMMs (dHMM) and black – continuous HMMs (cHMM). It is evident from Fig. 6.6 that our proposed representation has significantly higher mAP as compared to dHMMs and cHMMs while being consistent over the number of hidden states.

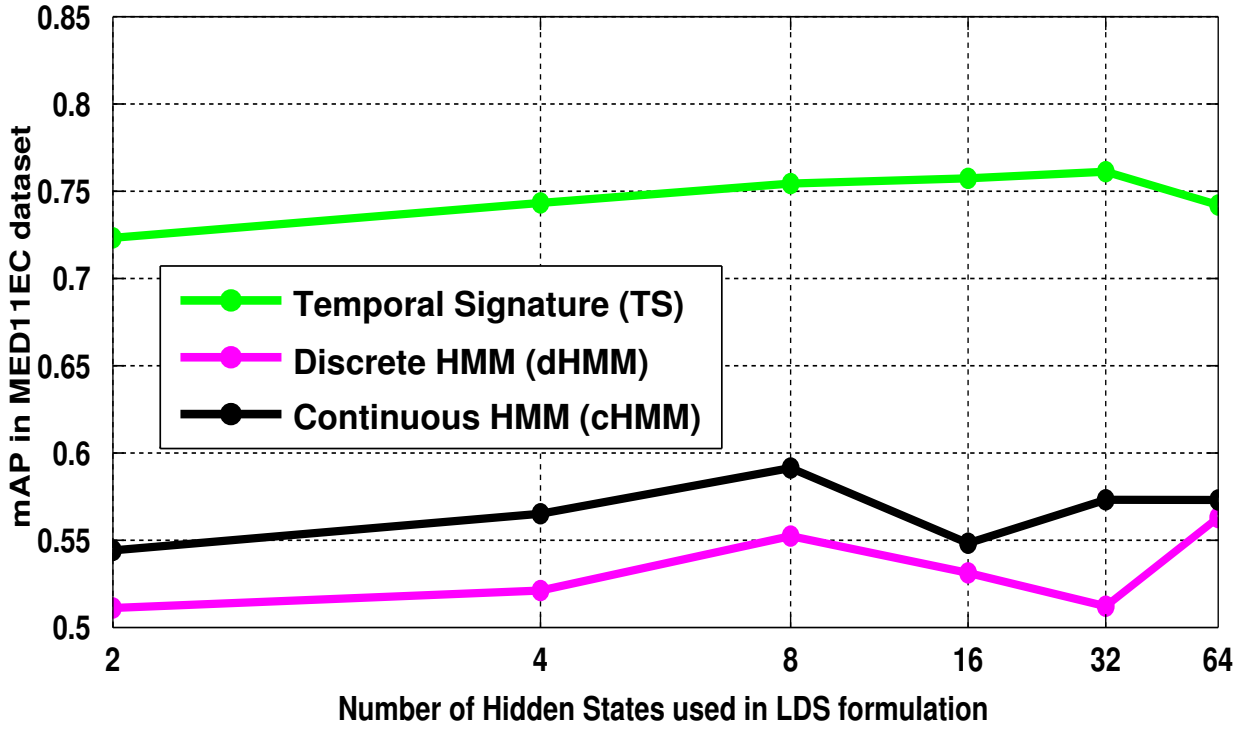


Figure 6.6: **Effect of hidden states on our LDS formulation**: mean Average Precision (mAP) scores of event detection in MED11EC is shown against number of hidden states ($\theta = 2, 4, 8, 16, 32, 64$) for learning LDS parameters using three similar approaches, with green – Temporal Signatures (TS), purple – discrete HMMs (dHMM) and black – continuous HMMs (cHMM). TS has significantly higher mAP scores compared to dHMMs and cHMMs while being consistent over the number of hidden states.

6.3.3 Parameter Selection for Temporal Features

In this section, we analyze the effect of different parameters involved in computation of the block Hankel matrix based descriptors and the temporal signatures. Fig. 6.7 empirically illustrates how different combination of parameters affect recognition performance over both MED11EC and MED12EC datasets. While constructing the block Hankel matrix based descriptors, we experiment with 3 different overlap settings, i.e. $r = 4, 8, 16$. Within each overlap setting, we use 4 different sets of largest eigenvectors ($m = 1, 2, 4, 8$) for the re-projection operation as described in Sect. 6.2.1, for computing the final descriptor. This gives an idea on the optimal number of auto-regressor coefficients in Eqn. (6.4) required to build an efficient representation. From Fig. 6.7, it is evident that selecting the optimal number of r and m does not significantly affect the performance of the proposed algorithm. This argues in favor of the robustness of the Hankel Matrix based descriptor. We observe that with an overlap setting of 16, and 8 largest eigenvectors yield slightly better than other settings, and hence use the same to report all our results. However, a lower overlap setting with fewer eigenvectors is advised, for additional computational benefits.

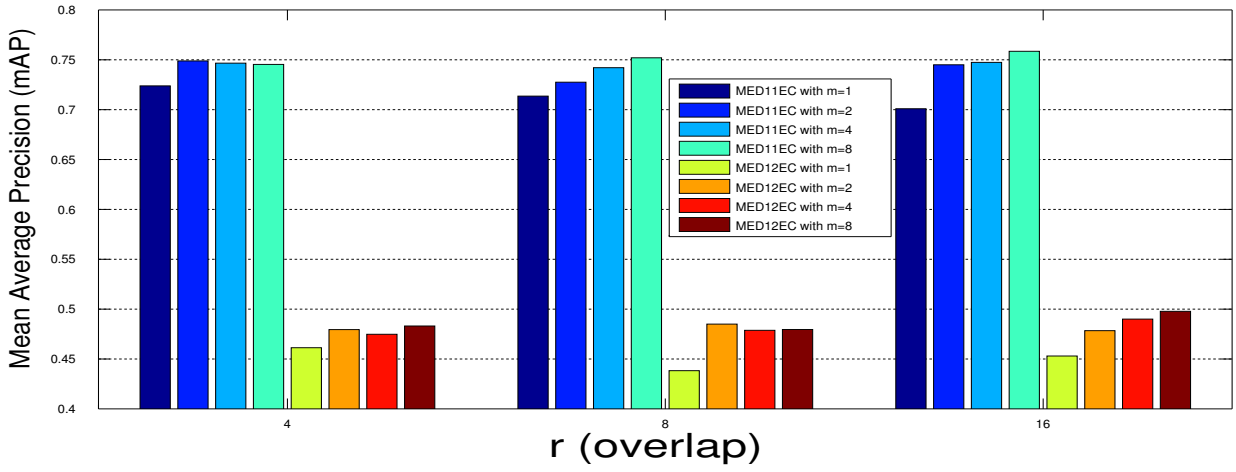


Figure 6.7: **Mean average precision on MED11EC and MED12EC using Hankel features with different overlap settings (r) and number of largest eigenvectors (m).**

Similar to the Hankel matrix based descriptors, we empirically determine the optimal length of the temporal signatures, required for effective recognition. The outcome of this experiment is presented in Fig. 6.8, where we vary the length of the temporal signature descriptors by altering the number of canonical coefficients in Eqn. (6.8). Similar to the experiments conducted on different Hankel Matrix based descriptor settings, we observe negligible change in mean average precision for both MED11EC and MED12EC datasets.

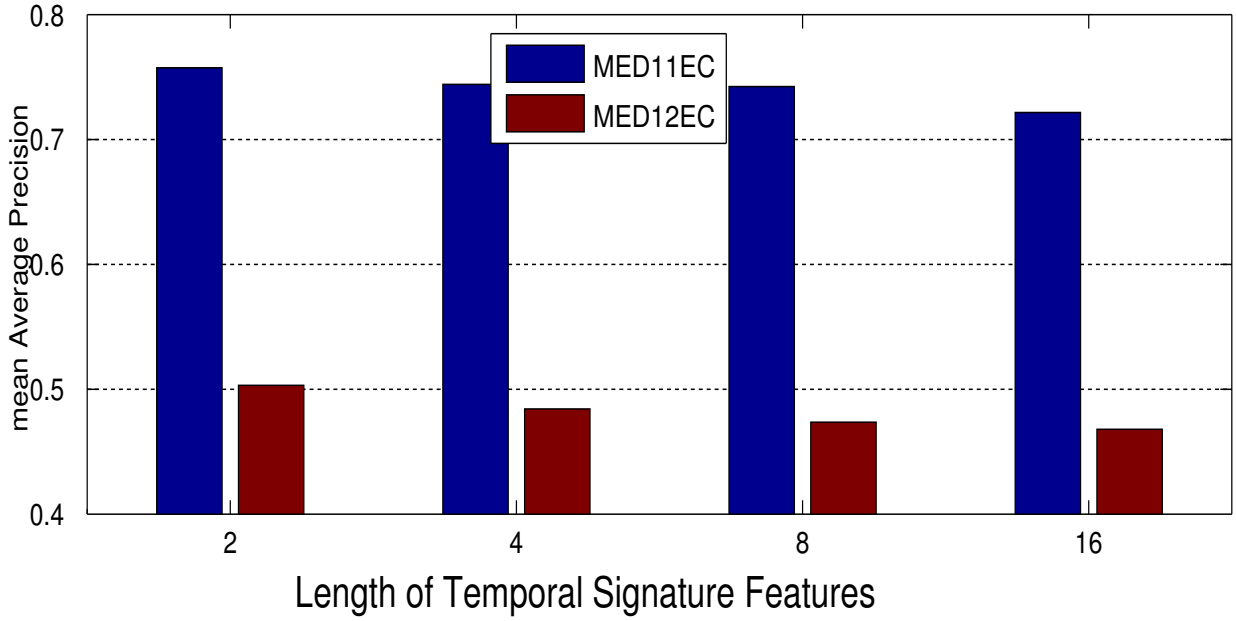


Figure 6.8: **Mean average precision on MED11EC and MED12EC using temporal signatures with different descriptor sizes.**

Furthermore, to show the benefits of combining both the proposed temporal representations, we fuse both the descriptors in a relatively simple early fusion framework. The fusion is performed using pure concatenation of the individually normalized descriptors. We report the performance in Fig. 6.9, Fig. 6.10 and Tab. 6.1. As evident in case of certain events e.g. *Getting a vehicle unstuck*, *Making a Sandwich*, *Winning a Race without a vehicle*, *Working on a Metal Crafts project*, *Attempting a Board Trick*, *Landing a Fish*, and *Renovating a Home* fusion increases the

mAP. To confirm, we perform a simple Fisher Discriminant analysis, for all fused descriptors of samples belonging to these classes, which depicts relatively larger ratios for inter-class distance versus intra-class scatter, when compared to other classes. In other words, the temporal features capture complementary temporal dynamics from the vector time series for the aforementioned event classes.

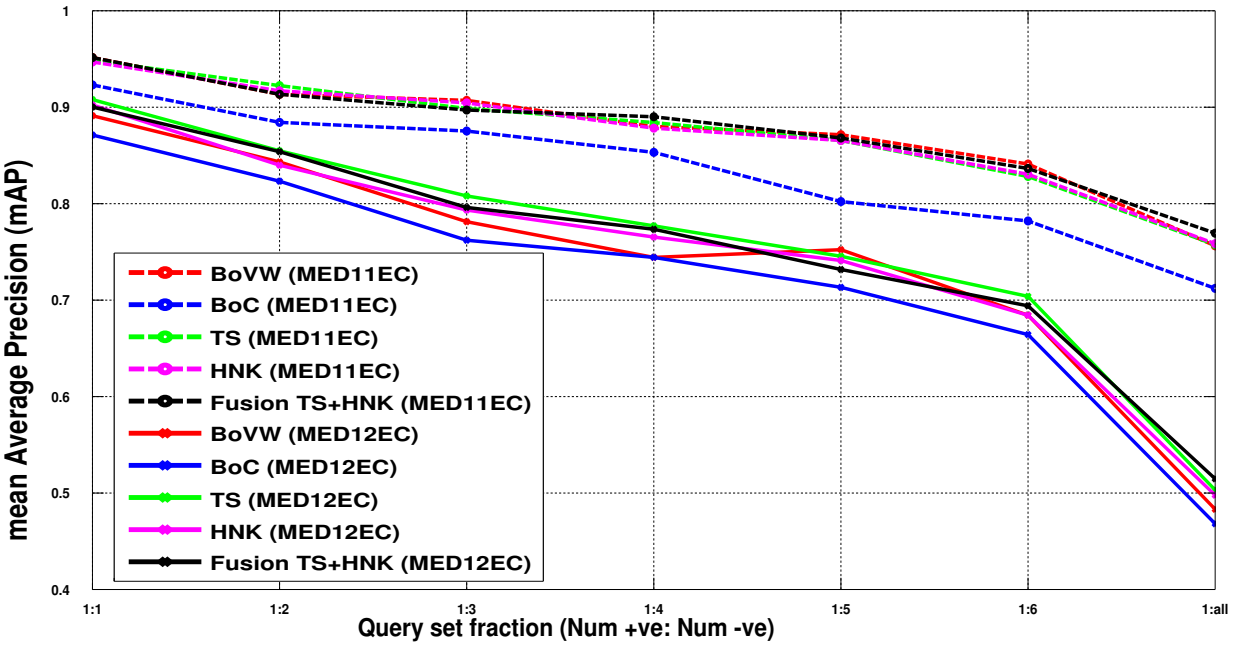


Figure 6.9: **mAP over different mixtures of query samples.**

Fig. 6.9 also provides an insight on the classifier's performance over different mixtures of query samples. Under this setting, we report the mAPs for both datasets, varying the number of positive samples and negative samples in the testing data. We begin with equal number of positive and negative samples, gradually increasing the number of negative samples keeping the number of positive samples fixed, until we exhausted all negative samples. Even with all the negative samples in the query set, we demonstrate respectable performance.

Although our approach is not directly comparable with purely flat-histogram based repre-

sentations, since these approaches ignore the additional temporal information, we provide comparative results on two different histogram based approaches. The first one is based on a bag-of-MBH-features representation computed on the videos with a visual vocabulary size of $2,048$ to make fair comparison with the concept representation vocabulary. A histogram intersection kernel is used in a binary SVM classifier to report the final event recognition performance This is listed as BoVW in Fig. 6.9. In the next setting, the vector time series flattened into a one-dimensional vector by averaging each vectors along time and is referred to BoC in 6.9. Interestingly, the bag-of-MBH-features based representation outperforms the naive concept based representation, which is also in accordance with previously reported results [67, 113]. However, it is worthwhile to observe that our temporal representation yields results comparable to the BoVW method, providing promising insights towards improving the overall event recognition performance when used in a hybrid fashion as suggested by [67]. Towards the end of Sect. 6.4, we confirm this hypothesis with conclusive experiments.

6.4 Results

In this section, we take the opportunity to report key results of our experiments. Tab. 6.1 reports the respective average precisions for individual event categories in MED11EC dataset. All the results reported here are on a query set consisting with a small number of positive samples for a given event category and all possible negative samples. Typically such a mixture consists of $20 - 50$ positive and $1,900$ negative samples. Hence the results reported here, can be easily generalized to results on the MED11 DEVT release by NIST, which contains a large number of videos not belonging to any of the pre-specified 15 event categories.

We begin comparing our proposed methods with the three different baseline methods mentioned previously, that also capture temporal information. It can be observed that features computed using DCT coefficients poorly represent our vector time series data. One reason for this

is the complex nature of our signals, which are not captured by the limited number of bases we use while computing the DCT based features. This reinforces the need for data-driven frequency grouping as proposed in Sect. 6.2.2.

Table 6.1: **Average Precision scores (MED11EC)** Performance of our proposed temporal features with contemporary methods that model temporal interactions on all 15 events in the MED11EC dataset.

	Avg. Prec. from Method					
Event	DCT	DHMM	CHMM	HNK	TS	CMB
E001	0.46	0.66	0.72	0.85	0.87	0.89
E002	0.44	0.64	0.71	0.89	0.91	0.91
E003	0.43	0.32	0.52	0.68	0.71	0.73
E004	0.39	0.39	0.39	0.61	0.59	0.59
E005	0.36	0.38	0.37	0.58	0.55	0.59
E006	0.34	0.38	0.51	0.87	0.87	0.85
E007	0.43	0.41	0.48	0.77	0.74	0.76
E008	0.67	0.69	0.71	0.88	0.89	0.87
E009	0.44	0.48	0.49	0.83	0.86	0.86
E010	0.38	0.48	0.51	0.74	0.75	0.75
E011	0.51	0.62	0.63	0.79	0.71	0.74
E012	0.37	0.73	0.68	0.78	0.76	0.78
E013	0.31	0.35	0.41	0.84	0.88	0.86
E014	0.34	0.46	0.48	0.68	0.68	0.67
E015	0.32	0.31	0.38	0.58	0.57	0.59
mAP	0.41	0.48	0.53	0.75	0.76	0.76

The next two columns in Tab. 6.1 reflect the results obtained after applying discrete and continuous HMM based classifiers to the discrete and continuous versions of the vector time-series data, respectively. Please refer to Sect. 6.3.2 on the discretization procedure. As there is no principled way to determine the optimal number of hidden states required by both the HMM based strategies, we resort to empirical techniques (Fig. 6.6), experimenting with different number of hidden states with corresponding Gaussian stochastic prior matrices. The discrete HMM based strategy is observed to perform best with 64 hidden states while the continuous HMM based strategy yields best mAP with merely 8 states. We conjecture that the nature of the distribution input to

both methods plays a major role in restricting the degrees of freedom of state transitions, thereby efficiently modeling the temporal structure with fewer parameters in case of continuous HMMs.

Conclusively, the continuous HMM based strategy outperforms the discrete version consistently in case of both MED11EC and MED12EC (Fig. 6.10). However, estimating the mixture parameters is a computationally intensive problem and a significant fraction of the videos in our datasets, the training does not converge, leading to fewer samples in training.

The last three columns in Tab. 6.1 report the respective mAP scores for the proposed representations starting with Hankel matrix based descriptors (HNK), followed by the Temporal Signatures (TS) and finally the combined feature obtained by early fusion of HNK and TS (CMB). Both of our proposed LDS based features perform better than the baseline methods by a significant margin ($22-35\%$), with temporal signatures (TS) slightly better than Hankel matrix based features (HNK). A broader comparison with all 25 events from MED12EC is provided in Fig. 6.10.

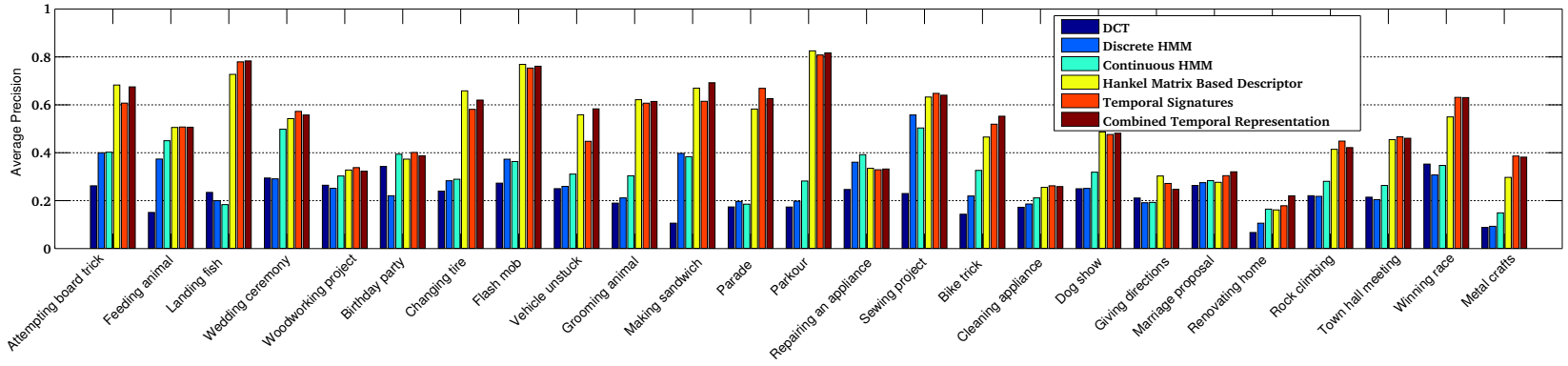


Figure 6.10: **Average Precision scores (MED12EC)** Performance of our proposed temporal features with contemporary methods that model temporal interactions for all 25 events in MED12EC.

We observe a significant reduction ($\sim 24\%$) in mAP when as we move to MED12EC from MED11EC (as seen in Fig. 6.9). This reduction can be attributed primarily due to the following reasons: firstly, the concepts defined in MED11EC do not have a significant overlap with the additional 10 events in MED12EC, as our concept vocabulary is constructed without keeping the new events under consideration. Consequently, there number of detectors required to model the whole MED12EC is not exhaustive. Secondly, there is a considerable change in the testing mixture when we move to MED12EC (from $20-50$+ve/$1,900$-ve to $20-50$+ve/$3,800$-ve). This results in some expected loss of performance for the same event categories in MED11EC and MED12EC.

Finally, in Tab. 6.2, we report additional experimental results to show how our improved representation of concepts, can be augmented with an existing low-level bag-of-features based representation to improve the overall mAP. We followed a late fusion based strategy to combine our classifier confidences with the one obtained from classifiers that operate on representation generated on top of purely low-level features. We use the same feature modality (MBH) to maintain consistency with results reported in Fig. 6.9.

For the sake of legibility, we list the mAPs obtained on MED11EC, and MED12EC using averaging concepts (BoC) over time, Bag of visual words (BoVW), and our early fusion of temporal features (CMB) in the first three columns of Tab. 6.2. The next two columns report late fusion of classifier confidences using BoC (Naive) and CMB (Final) with that from classifiers trained BoVW representations, respectively. As hypothesized, fusion of our combined temporal representation with BoVW yields superior performance as compared to regular averaging of concepts which has no notion of cross-concept temporal dependence. Although, the fusion results achieve only about $4-8\%$ performance gain over BoVW, it succeeds in making a stronger argument towards efficiently extracting temporal information from an intermediate representation without sacrificing the overall detection performance.

Table 6.2: **Fusion with BoVW** BoC, BoVW and CMB report the mAPs returned by classifiers training on representations constructed by averaging concepts over time, accumulating vector quantization and fusing proposed temporal descriptors, respectively. Naive and Final denote the results after employing late fusion of BoVW over BoC and BoVW over CMB, respectively.

	Fusion Methods				
Datasets	BoC	BoVW	CMB	Naive	Final
MED11EC	0.72	0.75	0.76	0.77	0.79
MED12EC	0.46	0.48	0.53	0.50	0.56

6.5 Discussions

In this section, we share some technical insights developed during the course of several experiments. It may occur to the interested reader that of all concepts we had enlisted and used for our experiments, how many are relevant towards the actual detection task? In order to answer this interesting question, we experiment with some of the state of the art automatic feature selection techniques [14, 28, 76, 95, 128, 188], well studied in machine learning literature. We integrate a subset of these techniques in our computational pipeline, just after the vector time series construction stage, with the goal of eliminating redundant concepts, thereby reducing feature computation overhead. Specifically, experiments are carried out using the Max-Relevance Min Redundancy (MRMR) [128] and RELIEF [76] feature selection algorithms. For both the algorithms, averaged 1-D versions of the vector time-series from different event categories are used as inputs. The algorithms return indices of relevant features based on mutual information content or other higher order statistics. This is followed by the regular temporal feature computation stage, which is performed over a dimensionally reduced vector series of only top 10 relevant concepts. Although this reduces the temporal feature computation complexity dramatically, it degrades the recognition performance by a significant extent.

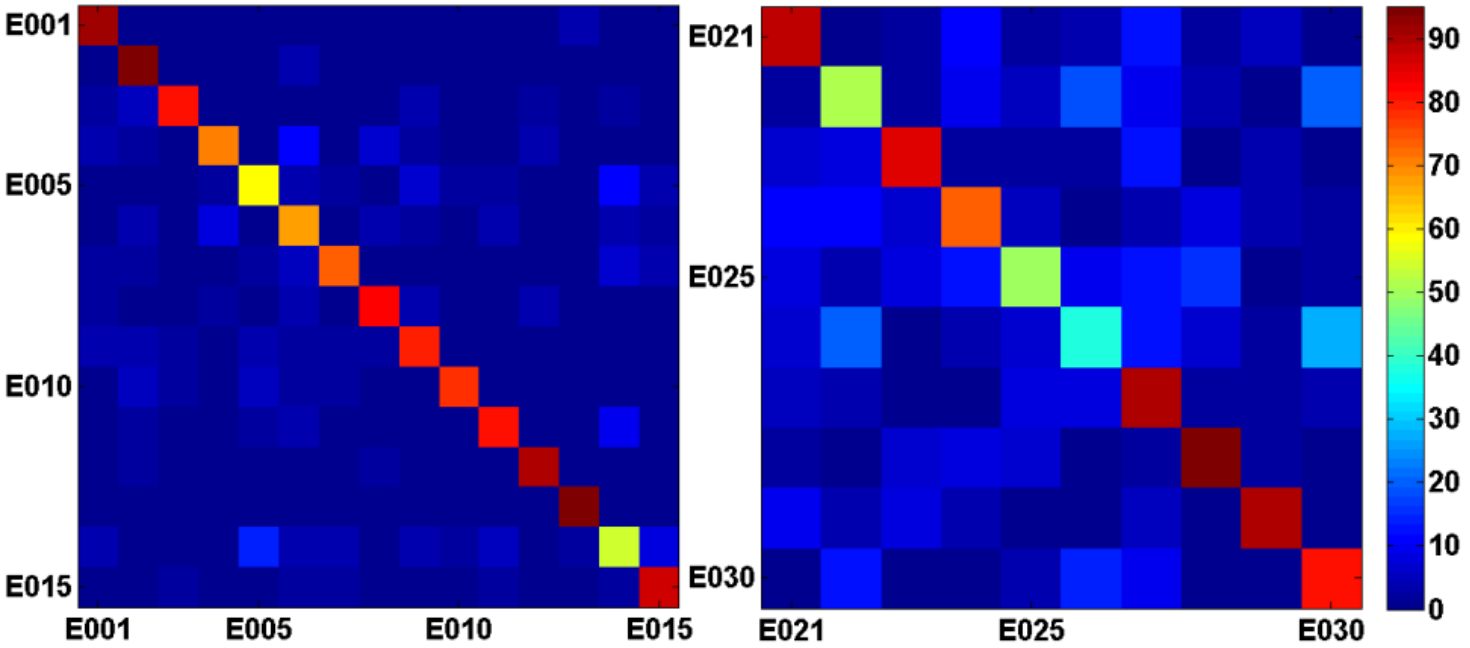


Figure 6.11: Confusion matrices obtained after using the optimal combinations of temporal features for EC1 (Left) and EC2 (right).

A detailed analysis of confusion encountered by classifiers can often be useful to design an optimal representation. This motivates us to perform a closed set multi-class classification within 15 events in MED11EC and 10 new events in MED12EC. Fig. 6.11, reports the respective confusion matrices for both event sets. The matrices provide an idea of the events that have similar temporal features owing to confusion in the recognition experiments. For example, the event "Attempting a board trick (E001)" is confused with "Parkour (E013)" as concept detectors for "jumping", "falling" etc. show similar temporal evolution. We also observe that "Renovating a home (E026)" is highly confused with "Working on a metal crafts project (E030)". This is because $\sim 60\%$ of their most confidently detected concepts share common evolutionary traits. This is a common trend across all new events in MED12EC.

6.6 Summary

Modeling temporal dynamics of spatio-temporal concepts occuring in a video can provide useful cue towards understanding the semantic structure of a video. We introduced two different techniques to model the temporal relationships between spatio-temporal concepts within the purview of a video using foundations from Linear dynamical systems. Through several in depth experiments, we demonstrated the efficacy of our proposed method over contemporary methods, that are used extensively by computer vision and multimedia researchers, to analyze temporal structure of videos. Although, our method does not significantly outperform bag-of-words based approaches, we believe, it can be used effectively for other relevant tasks such as multimedia event recounting which require better understanding of temporal structure present in multimedia data. As part of future work, we intend to extend this idea to a large corpus of concepts, which may be learned in an unsupervised fashion, given a collection of videos.

CHAPTER 7: FUTURE WORK

The previous chapters in this dissertation discusses a complete bottom up approach towards recognition of complex events in web videos. However, the visual recognition algorithms discussed so far, rely heavily on variants of supervised learning aproaches which require labeled training samples. Although these approaches have reported commendable success rates on benchmark datasets, their performance usually do not translate to larger datasets i.e. scale of YouTube or Flickr. This is primarily due to the assumption that both training and testing samples are drawn from the same distribution [17, 42, 53, 65, 103, 135, 142]. This causes well trained classifiers to fail miserably when subject to test samples that belong to domains different from source training domain. The term *domain* [142] in context of visual recognition can be interpreted as a closed setting defined by factors such as illumination, camera motion & orientation, background clutter, resolution etc. that directly affect the capture of an image or a video.

Since obtaining more annotated samples for each domain is not a practically feasible idea, there is a demanding requirement for efficient algorithms that are capable of propagating knowledge from existing representation into newer data. Inspring research [65, 103, 135] has been conducted in this direction, under the umbrella of visual *domain adaptation* or *transfer learning*. We intend to address this problem using foundations from deep convolutional neural networks (CNN) which have demonstrated exceptional credibility in recent large-scale object recognition challenges [79]. The success of CNNs can be attributed to their training which are inspired from mechanisms employed neural networks in biological vision. Similar to neurons, artificial CNNs exploit spatially local correlation by enforcing a local connectivity pattern between adjacent layers of receptive fields. However, simulating such sophisticated layered networks incur heavy computational complexity due to direct optimization of the supervised objective of interest.

In view of the above, two alternative approaches have been proposed namely — Restricted Boltzmann Machines (RBM) [16] and Denoising Auto-encoders (DA) [168, 169]. Both use a local

unsupervised criterion to pre-train each layer in order to produce a useful higher-level abstract representation at the current layer from the immediately lower layer. While RBMs are probabilistic models and require approximations which are usually intractable, DAs are deterministic and can be trained simply using gradient descent. Representations learned using layers of DAs (Stacked DAs) have been demonstrated to be effective in textual sentiment classification [33, 52] across different domains.

To the best of our knowledge, there has not been significant body of work involving SDAs that involve transfer learning to solve visual recognition tasks. Hence, we intend to demonstrate how SDAs can be used in context of learning useful representation from images across multiple visual domains [142]. Fig. 7.1 provides a bird-eye view of our proposed framework that can be used to learn meaningful representations across different domains, in order to achieve robust recognition performance.

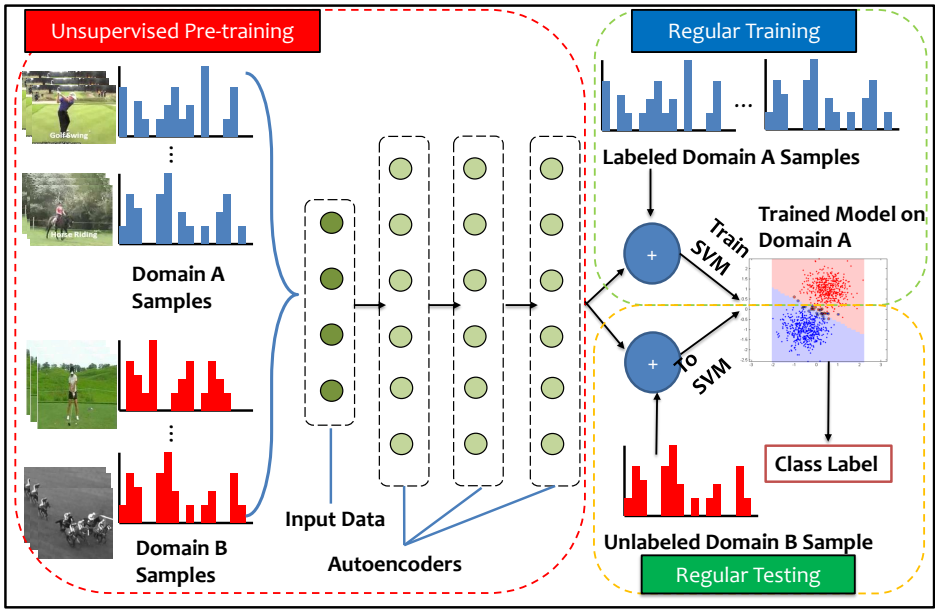


Figure 7.1: **Various stages involved in training a two layered SDAs from two domains**: Data from both input domains are corrupted and mapped to a combined hidden representation using SDA (output). Next labeled samples from domain A and unlabeled samples from domain B are combined with mapped output which form the final representation to train a classifier (linear SVM) and subsequently test samples from domain B.

We conjecture that the methodology proposed in Fig. 7.1 can be adapted to address challenging action recognition problems across multiple views [179] and across datasets [22, 29]. In addition, we can build on the experience on objects and actions, and apply this approach into recognition of complex events in scenarios where very few or no training samples are available for target events. We envision that this fundamentally different perspective to cross-domain transfer learning using SDAs would facilitate solving a broader spectrum of computer vision problems. We envision that this methodology, in conjunction with the previous set of work discussed in Chapters 3 , 4, 5, and 6 can eventually be used to close the loop in recognition of complex events from unconstrained open-source Internet videos.

CHAPTER 8: CONCLUSIONS

In this dissertation we presented a set of methodologies with an end objective to perform semantic analysis of open-source consumer uploaded Internet videos depicting complex events. In our approach, we first presented a principled decomposition of complex events into hierarchical components and performed an in-depth analysis of how existing research are being used to cater to various levels of this hierarchy. Next we identified three key stages where we highlighted our technical contributions, emphasizing on a recognition framework which is more semantically driven. As part of the first stage, we introduced two novel semi-global features which can be used to capture complementary information from videos.

In this context, we introduced a video-level feature that encapsulates coarse statistics pertaining to camera motion present in typical consumer uploaded videos. We devised this novel feature on top of inter-frame homographies which using *Lie algebra of projective groups*, are transformed to an intermediate vector space that preserves the intrinsic geometric structure of the transformation. Multiple time series are then constructed from these mappings. Final features computed from time series are used for discriminative classification of video shots. Additionally, we demonstrated how this feature can be used as a source of complementary information for recognition of complex events in videos. We also observed that global spatio-temporal context plays an important role in analysis of web videos, which motivated us to propose compact clip level descriptors for such videos based on *covariance of low-level appearance and motion features*. These features were later assimilated into a sparse coding framework to recognize realistic actions and gestures. Within this, the sparse approximation of a set of covariance matrices is treated as a *determinant maximization* problem where the bases (covariance matrices) are obtained from training videos. We evaluated the proposed technique with a sparse linear approximation alternative suitable for equivalent *vector spaces of covariance matrices* using *Orthogonal Matching Pursuit*. A variety of experimental settings validates our hypothesis that contextual information in the form

of camera motion, background motion and even correlation between different individual feature modalities are vital cues for the analysis of web videos.

In the next chapter of this dissertation we switched our focus from low-level or semi-global features to intermediate representations which in practice, is an equally important area of research in video understanding. We presented an efficient probabilistic alternative to the traditional bag-of-features based representation from low-level features computed from videos. Since we iteratively generate a *Maximum Likelihood estimate* of an instance given a set of characteristic features which can be sampled randomly, our representation is conceptually more elegant and computationally superior to the quantization approaches used in traditional bag of features based techniques. In addition to commendable performance in standard action and scene recognition datasets, we demonstrate substantial improvement in detection of large scale semantically accurate, human-understandable mid-level spatio-temporal concepts for modeling complex events.

In the concluding technical section of this dissertation, we insinuated two discriminative feature spaces to model temporal interactions between spatio-temporal concepts which can be efficiently integrated into existing classifiers for complex event recognition. The first in these lines is based on *Subspace state identification techniques*, wherein *Block Hankel Matrices* are constructed from temporally evolving sequences of concept detector responses, followed by their Eigen decomposition to yield compact descriptors. The second exploits statistically meaningful characteristics from multiple interacting time-series such as lag-independence, harmonics, frequency proximity etc., grouping similar temporally evolving processes into identical *harmonic groups* using an expectation minmization algorithm. Through thorough experiments, we exhibited state of the art performance in complex event recognition on benchmark TRECVID MED 11-12 datasets.

Although the suggested approaches address major issues in complex event recognition from multiple perspectives, their success for some of the frequently encountered issues - such as recognizing events with very few training samples, is not conclusive. In future, we intend to address such issues leveraging on information available from external sources such as web corpus, social media,

and other image/video sharing sites, using advanced domain transfer techniques. We provided a potential insight along this direction in our future work. We conjecture that this dissertation will provide a significant amount of knowledge to researchers and practitioners in both computer vision and multimedia communities, who are interested in solving the challenging problem of complex event recognition.

LIST OF REFERENCES

[1] NIST TRECVID Multimedia Event Detection (MED) task. http://www.nist.gov/itl/iad/mig/med.cfm.

[2] A framework for flexible summarization of racquet sports video using multiple modalities. *CVIU*, 113(3):415 – 424, 2009.

[3] University of Central Florida 50 human action dataset. http://server.cs.ucf.edu/~vision/data/UCF50.rar, 2010.

[4] YouTube Statistics. http://www.youtube.com/t/press_statistics, Jan. 2013.

[5] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3), 2011.

[6] S. Ali, A. Basharat, and M. Shah. Chaotic invariants for human action recognition. In *Proc. of IEEE International Conference on Computer Vision*, 2007.

[7] S. Ali and M. Shah. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):288–303, 2010.

[8] A. Andoni and P. Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Communications of the ACM*, 51(1), 2008.

[9] D. Arijon. *Grammar of the Film Language*. Hasting House Publishers, NY.

[10] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2):411–421, aug 2006.

[11] M. Baillie and J. M. Jose. Audio-based event detection for sports video. In *Proc. International Conference on Image and Video Retrieval*, 2003.

[12] L. Ballan, M. Bertini, A. D. Bimbo, L. Seidenari, and G. Serra. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications*, 51(1):279–302, 2011.

[13] A. Barbu, A. Bridge, D. Coroian, S. Dickinson, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang. Large-scale automatic labeling of video events with verbs based on event-participant interaction. In *arXiv:1204.3616v1*, 4 2012.

[14] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Networks*, 5:537–550, 1994.

[15] H. Bay, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. In *Proc. of European conference on Computer vision*, pages 404–417, 2006.

[16] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Proc. Neural Information Processing Systems*, 2007.

[17] A. Bergamo and L. Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Proc. Neural Information Processing Systems*, pages 181–189, 2010.

[18] S. Bhattacharya, M. Kaleyeh, R. Sukthankar, and M. Shah. Understanding temporal dynamics of low-level concepts for complex event recognition. In *Proc. of ACM Multimedia (under review)*, 2013.

[19] S. Bhattacharya, R. Mehran, R. Sukthankar, and M. Shah. Cinematographic shot classification and its application to complex event recognition. *IEEE Trans. MM (under review)*, 2012.

[20] S. Bhattacharya, N. Souly, and M. Shah. Covariance of motion and appearance features for spatio temporal recognition tasks. *IEEE Trans. PAMI (under review)*, 2012.

[21] S. Bhattacharya, R. Sukthankar, R. Jin, and M. Shah. A probabilistic representation for efficient large scale visual recognition tasks. In *Proc. IEEE CVPR (CVPR)*, pages 2593–2600, 2011.

[22] W. Bian, D. Tao, and Y. Rui. Cross-domain human action recognition. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 42(2):298–307, 2012.

[23] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *Proc. of IEEE International Conference on Computer Vision*, pages 1395–1402, 2005.

[24] A. F. Bobick. Movement, activity, and action: The role of knowledge in the perception of motion. 352:1257–1265, 1997.

[25] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[26] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30, 2007.

[27] D. Brezeale and D. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(3):416–430, 2008.

[28] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *J. Mach. Learn. Res.*, 13:27–66, Mar. 2012.

[29] L. Cao, Z. Liu, and T. Huang. Cross-dataset action detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1998–2005, 2010.

[30] L. Cao, Y. Mu, A. Natsev, S.-F. Chang, G. Hua, and J. R. Smith. Scene aligned pooling for complex video recognition. In *Proc. European Conference on Computer Vision*, pages 688–701, 2012.

[31] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[32] C.-Y. Chen and K. Grauman. Watching unlabeled video helps learn new human actions from very few labeled snapshots. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[33] M. Chen, Z. E. Xu, K. Q. Weinberger, and F. Sha. Marginalized denoising autoencoders for domain adaptation. In *Proc. of International Conference on Machine Learning*, 2012.

[34] H. Cheng, A. Tamrakar, S. Ali, Q. Yu, O. Javed, J. Liu, A. Divakaran, H. S, A. Hauptmann, M. Shah, S. Bhattacharya, M. Witbrock, J. Curtis, R. Mertens, T. Darrell, R. Manmatha, and J. Allan. Team sri-sarnoff's aurora system @ trecvid 2011. In *Proc. NIST TRECVID Workshop*, 2011.

[35] C. I. Connolly. Learning to Recognize Complex Actions using Conditional Random Fields. In *Proceedings of the 3rd international conference on Advances in visual computing - Volume Part II*, ISVC'07, pages 340–348, 2007.

[36] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *Proc. of European conference on Computer vision*, 2004.

[37] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

[38] C. DiMonte and K. Arun. Tracking the frequencies of superimposed time-varying harmonics. In *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pages 2539–2542, apr 1990.

[39] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. 2005. IEEE International Workshop on VS-PETS.

[40] A. Doulamis and N. Doulamis. Optimal content-based video decomposition for interactive video navigation. *Circuits and Systems for Video Technology, IEEE Transactions on*, 14(6):757 – 775, june 2004.

[41] T. Drummond and R. Cipolla. Application of lie algebras to visual servoing. *International Journal of Computer Vision*, 37:2000, 1999.

[42] L. Duan, D. Xu, and S. F. Chang. Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1338–1345, 2012.

[43] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99, 2011.

[44] R. Fablet, P. Bouthemy, and P. Perez. Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval. *IEEE Transactions on Image Processing*, 11:393–407, 2002.

[45] J. Fan and H. Luo. Principal video shot: Linking low-level perceptional features to semantic video events. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, volume 4, page 37, june 2003.

[46] R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training support vector machines. *J. Mach. Learn. Res.*, 6:1889–1918, 2005.

[47] J. Farquhar, S. Szedmak, H. Meng, and J. Shawe-Taylor. Improving "bag-of-keypoints" image categorisation. Technical report, University of Southampton, 2005.

[48] A. Fathi and G. Mori. Action recognition by learning mid-level motion features. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[49] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[50] W. Forstner and B. Moonen. A metric for covariance matrices. In *TR Dept. of Geodesy and Geoinformatics, Stuttgart University*, 1999.

[51] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky. Hough forests for object detection, tracking, and action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(11):2188–2202, 2011.

[52] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proc. of International Conference on Machine Learning*, pages 513–520, 2011.

[53] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Proc. IEEE International Conference on Computer Vision*, pages 999–1006, 2011.

[54] K. Grauman and T. Darrell. The pyramid match kernel: discriminative classification with sets of image features. In *Proc. of IEEE International Conference on Computer Vision*, 2005.

[55] K. Guo, P. Ishwar, and J. Konrad. Action recognition in video by sparse representation on covariance manifolds of silhouette tunnels. In *Proceedings of the 20th International conference on Recognizing patterns in signals, speech, images, and videos*, ICPR'10, pages 294–305, 2010.

[56] K. Guo, P. Ishwar, and J. Konrad. Action recognition using sparse representation on covariance manifolds of optical flow. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 188–195, 2010.

[57] Y. Habibolu, O. Gnay, and A. etin. Covariance matrix-based fire and flame detection method in video. *Machine Vision and Applications*, pages 1–11, September 2011.

[58] T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18:607–616, June 1996.

[59] S. Hongeng, R. Nevatia, and F. Bremond. Video-based Event Recognition: Activity Representation and Probabilistic Recognition Methods. *CVIU*, 96(2):129–162, 2004.

[60] C.-L. Huang, H.-C. Shih, and C.-Y. Chao. Semantic Analysis of Soccer Videos using Dynamic Bayesian Network. *IEEE Transactions on Multimedia*, 8(4):749 –760, 2006.

[61] N. Inoue, Y. Kamishima, T. Wada, K. Shinoda, and S. Sato. TokyoTech+Canon at TRECVID 2011. In *Proc. NIST TRECVID Workshop*, 2011.

[62] S. S. Intille and A. F. Bobick. Recognizing Planned, Multiperson Action. *CVIU*, 81(3):414 – 445, 2001.

[63] E. A. Jackson. *Perspectives of Nonlinear Dynamics, Chapter 2*. Cambridge University Press, 1991.

[64] W. James and J. Stein. Estimation with Quadratic Loss. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, pages 361–379, 1961.

[65] I.-H. Jhuo, D. Liu, D. T. Lee, and S. F. Chang. Robust visual domain adaptation with low-rank reconstruction. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[66] S. Ji, W. Xu, M. Yang, and K. Yu. 3d convolutional neural networks for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):221–231, 2013.

[67] Y. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S. Chang. Columbia-UCF TRECVID 2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching. *Proc. of NIST TRECVID Workshop*, Dec. 2010.

[68] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, and M. Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, November 2012.

[69] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. 2007.

[70] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann. Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Trans. Multimedia*, 12(1):42–53, 2010.

[71] I. Junejo, E. Dexter, I. Laptev, and P. Perez. View-independent action recognition from temporal self-similarities. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(1):172–185, 2011.

[72] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *Proc. of IEEE International Conference on Computer Vision*, 2005.

[73] Y. Ke, R. Sukthankar, and M. Hebert. Efficient visual event detection using volumetric features. In *Proc. IEEE International Conference on Computer Vision*, volume 1, pages 166–173 Vol. 1, 2005.

[74] Y. Ke, R. Sukthankar, and M. Hebert. Event detection in crowded videos. In *Proc. IEEE International Conference on Computer Vision*, pages 1–8, 2007.

[75] Y. Ke, R. Sukthankar, and M. Hebert. Spatio-temporal shape and flow correlation for action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[76] K. Kira and L. A. Rendell. A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning*, ML92, pages 249–256, 1992.

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

[78] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2046–2053, 2010.

[79] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Neural Information Processing Systems*, pages 1106–1114, 2012.

[80] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proc. of IEEE International Conference on Computer Vision*, 2011.

[81] I. Laptev and T. Lindeberg. Space-time interest points. In *Proc. of IEEE International Conference on Computer Vision*, pages 432–439, 2003.

[82] G. Lavee, E. Rivlin, and M. Rudzsky. Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in videos. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 39(5):489–504, 2009.

[83] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[84] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3361–3368, 2011.

[85] K. Lee and D. P. W. Ellis. Audio-based semantic concept classification for consumer video. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1406–1416, 2010.

[86] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001.

[87] B. Li, M. Ayazoglu, T. Mao, O. I. Camps, and M. Sznaier. Activity recognition using dynamic subspace angles. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3200, 2011.

[88] B. Li, O. I. Camps, and M. Sznaier. Cross-view Activity Recognition using Hankelets. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, june 2012.

[89] L. Li, B. A. Prakash, and C. Faloutsos. Parsimonious Linear Fingerprinting for Time-series. *Proc. VLDB Endow.*, 3(1-2):385–396, Sept. 2010.

[90] P. Li and Q. Sun. Tensor-based covariance matrices for object tracking. In *Proc. IEEE International Conference on Computer Vision Workshops*, pages 1681–1688, 2011.

[91] R. Li and R. Chellappa. Group motion segmentation using a spatio-temporal driving force model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2038–2045, 2010.

[92] W. Li, Z. Zhang, and Z. Liu. Expandable Data-Driven Graphical Modeling of Human Actions Based on Salient Postures. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(11):1499–1510, 2008.

[93] D. Lin, E. Grimson, and J. W. F. III. Modeling and estimating persistent motion with geometric flows. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2010.

[94] D. Lin, W. E. L. Grimson, and J. W. F. III. Learning visual flows: A lie algebraic approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 747–754, 2009.

[95] D. Lin and X. Tang. Conditional infomax learning: An integrated framework for feature extraction and fusion. In *Proc. European Conference on Computer Vision*, pages 68–82, 2006.

[96] Z. Lin, Z. Jiang, and L. S. Davis. Recognizing actions by shape-motion prototype trees. In *Proc. of IEEE International Conference on Computer Vision*, 2009.

[97] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3337–3344, 2011.

[98] J. Liu and M. Shah. Scene modeling using co-clustering. In *Proc. of IEEE International Conference on Computer Vision*, 2007.

[99] J. Liu and M. Shah. Learning human action via information maximization. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[100] J. Liu, Y.Yang, and M. Shah. Learning semantic visual vocabularies using diffusion distance. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[101] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2), 2004.

[102] Y. M. Lui, J. Beveridge, and M. Kirby. Action classification on product manifolds. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 833 –839, 2010.

[103] J. Luo, T. Tommasi, and B. Caputo. Multiclass transfer learning from unconstrained priors. In *Proc. IEEE International Conference on Computer Vision*, pages 1863–1870, 2011.

[104] R. F. Lyon, M. Rehn, S. Bengio, T. C. Walters, and G. Chechik. Sound retrieval and ranking using sparse auditory representations. *Neural Computation*, 22(9):2390–2416, 2010.

[105] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 –8, june 2008.

[106] M. Marszałek, I. Laptev, and C. Schmid. Actions in context. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[107] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The det curve in assessment of detection task performance. In *European Conference on Speech Communication and Technology*, pages 1895–1898, 1997.

[108] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Proc. Neural Information Processing Systems*, pages 985–992, 2006.

[109] N. Morsillo, G. Mann, and C. Pal. Youtube scale, large vocabulary video annotation. *Chapter 14 in Video Search and Mining, Springer-Verlag series on Studies in Computational Intelligence*, pages 357–386, 2010.

[110] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *International Conference on Computer Vision Theory and Application VISSAPP'09)*, pages 331–340. INSTICC Press, 2009.

[111] J. Nam, M. Alghoniemy, and A. H. Tewfik. Audio-visual content-based violent scene characterization. In *in IEEE International Conference on Image Processing*, pages 353–357, 1998.

[112] M. Naphade, J. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 13(3):86 –91, 2006.

[113] P. Natarajan, P. Natarajan, V. Manohar, S. Wu, S. Tsakalidis, S. N. Vitaladevuni, X. Zhuang, and R. Prasad. BBN VISER TRECVID 2011 multimedia event detection system. In *Proc. NIST TRECVID Workshop*, 2011.

[114] P. Natarajan and R. Nevatia. Online, Real-time Tracking and Recognition of Human Actions. In *Proc. IEEE Workshop MVC*, pages 1–8, 2008.

[115] A. Natsev, J. R. Smith, M. Hill, G. Hua, B. Huang, M. Merler, L. Xie, H. Ouyang, and M. Zhou. IBM Research TRECVID-2010 video copy detection and multimedia event detection system. In *Proc. NIST TRECVID Workshop*, 2010.

[116] C.-W. Ngo, T.-C. Pong, and H. Zhang. On clustering and retrieval of video shots through temporal slices analysis. *IEEE Transactions on Multimedia*, pages 446–458, 2002.

[117] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang. Motion-based video representation for scene change detection. *Int. J. Comput. Vision*, 50(2):127–142, Nov. 2002.

[118] C.-W. Ngo, T.-C. Pong, and H.-J. Zhang. Motion analysis and segmentation through spatio-temporal slices processing. *IEEE Transactions on Image Processing*, 12:341–355, 2003.

[119] J. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proc. of European conference on Computer vision*, volume 6312 of *Lecture Notes in Computer Science*, pages 392–405. 2010.

[120] J. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[121] J. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, 79(3):299–318, 2008.

[122] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. Roy-Chowdhury, and M. Desai. A large-scale benchmark dataset for event recognition in surveillance video. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[123] A. Oliva and A. Torralba. Building the gist of a scene: the role of global image features in recognition. In *Progress in Brain Research*, 2006.

[124] P. V. Overschee and B. D. Moor. N4SID: Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems. *Automatica*, 30(1):75–93, 1994.

[125] Y. Pang, Y. Yuan, and X. Li. Gabor-based region covariance matrices for face recognition. *IEEE Trans. Circuits Syst. Video Techn.*, 18(7), 2008.

[126] S.-C. Park, H.-S. Lee, and S.-W. Lee. Qualitative estimation of camera motion parameters from the linear composition of optical flow. *Pattern Recognition*, 37(4):767–779, 2004.

[127] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In *Proc. International Symposium on Hearing*, pages 429–446, 1992.

[128] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:1226–1238, 2005.

[129] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[130] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *Proc. of European conference on Computer vision*, 2006.

[131] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[132] R. Poppe. Survey on vision-based human action recognition. *Image and Vision Computing*, 28(6):976–990, 2010.

[133] F. Porikli and T. Kocak. Robust license plate detection using covariance descriptor in a neural network framework. In *Proceedings of the IEEE International Conference on Video and Signal Based Surveillance*, AVSS '06, pages 107–, 2006.

[134] F. Porikli, O. Tuzel, and P. Meer. Covariance tracking using model update based on lie algebra. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2006.

[135] G.-J. Qi, C. Aggarwal, Y. Rui, Q. Tian, S. Chang, and T. Huang. Towards cross-category knowledge propagation for learning visual concepts. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 897–904, 2011.

[136] Y. Qi, A. Hauptmann, and T. Liu. Supervised classification for video shot segmentation. In *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pages 689–692, 2003.

[137] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Trans. on Circuits and Systems for Video Technology*, 15:52–64, 2003.

[138] A. Ravichandran and R. Vidal. Video registration using dynamic textures. *IEEE Trans. PAMI*, 33(1):158–171, 2011.

[139] M. J. Roach, J. Mason, M. Pawlewski, M. Heath, and I. I. Re. Video genre classification using dynamics. In *ICASSP*, 2001.

[140] R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. Technical report, Computer Science Department, Technion (Israel Institute of Technology), 2008.

[141] S. Sadanand and J. J. Corso. Action bank: A high-level representation of activity in video. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1234–1241, 2012.

[142] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. European Conference on Computer Vision*, pages 213–226, 2010.

[143] S. Savarese, A. DelPozo, J. Niebles, and L. Fei-Fei. Spatial-temporal correlatons for unsupervised action classification. In *IEEE Workshop on Motion and video Computing*, pages 1–8, 2008.

[144] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.

[145] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of ACM MM*, 2007.

[146] R. Sivalingam, D. Boley, V. Morellas, and N. Papanikolopoulos. Tensor sparse coding for region covariances. In *Proceedings of the 11th European conference on Computer vision: Part IV*, pages 722–735, 2010.

[147] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In *Proc. of IEEE International Conference on Computer Vision*, 2003.

[148] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. ACM International Workshop on Multimedia Information Retrieval*, 2006.

[149] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[150] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006.

[151] C. G. M. Snoek and M. Worring. Concept-based video retrieval. *Foundations and Trends in Information Retrieval*, 2(4):215–322, 2008.

[152] Y. Song, M. Zhao, J. Yagnik, and X. Wu. Taxonomic classification for web-based videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 871–878, 2010.

[153] M. V. Srinivasan, S. Venkatesh, and R. Hosie. Qualitative estimation of camera motion parameters from video sequences. *Pattern Recognition*, 30(4):593–606, 1997.

[154] T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *In Intl. Workshop on Automatic Face and Gesture Recognition*, 1995.

[155] X. Sun, M. Chen, and A. Hauptmann. Action recognition via local descriptors and holistic features. In *Computer Vision and Pattern Recognition Workshop*, 2009.

[156] S. Takagi, S. Hattori, K. Yokoyama, A. Kodate, and H. Tominaga. Sports video categorizing method using camera motion parameters. In *in Proceedings of ICME*, pages 461–464, 2003.

[157] F. Takens. Detecting strange attractors in turbulence. *Journal of Discrete Algorithms*, 1981.

[158] A. Tamrakar et al. Evaluation of Low-level Features and their Combinations for Complex Event Detection in Open Source Videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 3681–3688, june 2012.

[159] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655 –4666, dec. 2007.

[160] P. K. Turaga and R. Chellappa. Locally time-invariant models of human activities using trajectories on the grassmannian. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2435–2441, 2009.

[161] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488, 2008.

[162] T. Tuytelaars and C. Schmid. Vector quantizing feature space with a regular lattice. In *Proc. of IEEE International Conference on Computer Vision*, 2007.

[163] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Proc. of European conference on Computer vision*, pages 589–600, 2006.

[164] O. Tuzel, F. Porikli, and P. Meer. Human detection via classification on riemannian manifolds. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[165] D. L. Vail, M. M. Veloso, and J. D. Lafferty. Conditional Random Fields for Activity Recognition. In *Proc. IJCAIMS*, 2007.

[166] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[167] J. Van Gemert, J. Geusebroek, J. Veenman, and A. Smeulders. Kernel codebooks for scene categorization. In *Proc. of European conference on Computer vision*, 2008.

[168] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proc. of International Conference on Machine Learning*, pages 1096–1103, 2008.

[169] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, Dec. 2010.

[170] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin. Action recognition by dense trajectories. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[171] H. L. Wang and L.-F. Cheong. Taxonomy of directing semantics for film shot classification. *IEEE Trans. Cir. and Sys. for Video Technol.*, 19, 2009.

[172] L. Wang and D. Suter. Recognizing Human Activities from Silhouettes: Motion Subspace and Factorial Discriminative Graphical Model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[173] S. Wang, S. Jiang, Q. Huang, and W. Gao. Shot classification for action movies based on motion characteristics. In *International Conference on Image Processing*, pages 2508–2511, 2008.

[174] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. H. Tu. Shape and appearance context modeling. In *Proc. of IEEE International Conference on Computer Vision*, 2007.

[175] Y. Wang, K. Huang, and T. Tan. Group Activity Recognition Based on ARMA Shape Sequence Modeling. volume 3, pages 209–212, 2007.

[176] Y. Wang and G. Mori. Max-margin hidden conditional random fields for human action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 872–879, 2009.

[177] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. YouTubeCat: Learning to categorize wild web videos. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 879–886, 2010.

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

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

[180] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proc. of European conference on Computer vision*, 2008.

[181] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. D. Tardos. An image-to-map loop closing method for monocular slam. In *Proc. International Conference on Intelligent Robots and and Systems*, 2008.

[182] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proc. of IEEE International Conference on Computer Vision*, 2005.

[183] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[184] L. Xie, P. Xu, S.-F. Chang, A. Divakaran, and H. Sun. Structure Analysis of Soccer Video with Domain Knowledge and Hidden Markov Models. *Pattern Recognition Letters*, 25(7):767–775, 2004.

[185] M. Xu, N. C. Maddage, C. Xu, M. Kankanhalli, and Q. Tian. Creating audio keywords for event detection in soccer video. In *Proc. IEEE International Conference on Multimedia and Expo*, 2003.

[186] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, 1992.

[187] J. Yamato, J. Ohya, and K. Ishii. Recognizing Human Action in Time-Sequential Images using Hidden Markov Model. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1992.

[188] H. H. Yang and J. E. Moody. Data visualization and feature selection: New algorithms for nongaussian data. In *Proc. Neural Information Processing Systems*, pages 687–702, 1999.

[189] Y. Yang and M. Shah. Complex events detection using data-driven concepts. In *Proc. European Conference on Computer Vision*, ECCV'12, pages 722–735, 2012.

[190] A. Yao, J. Gall, and L. Van Gool. A hough transform-based voting framework for action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 2061–2068, 2010.

[191] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proc. IEEE International Conference on Computer Vision*, pages 1331–1338, 2011.

[192] J. Yao and J.-M. Odobez. Fast human detection from videos using covariance features. In *The Eighth International Workshop on Visual Surveillance - VS2008*, 2008.

[193] Y. Zhai, J. Liu, X. Cao, A. Basharat, A. Hakeem, S. Ali, M. Shah, C. Grana, and R. Cucchiara. Video understanding and content-based retrieval. In *In Proceedings of NIST TREC Video Retrieval*, 2005.

[194] Z. Zhang and D. Tao. Slow feature analysis for human action recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(3):436–450, 2012.

[195] X. Zhu, X. Xue, J. Fan, and L. Wu. Qualitative camera motion classification for content-based video indexing. In *Advances in Multimedia Information Processing*, pages 1128–1136, 2002.