

LEVERAGING CONTEXTUAL CUES FOR DYNAMIC SCENE UNDERSTANDING

A Thesis
Presented to
The Academic Faculty

by

Vinay Bettadapura

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Interactive Computing

Georgia Institute of Technology
May 2016

Copyright © 2016 by Vinay Bettadapura

LEVERAGING CONTEXTUAL CUES FOR DYNAMIC SCENE UNDERSTANDING

Approved by:

Professor Irfan Essa, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Rahul Sukthankar
Google Research
Google Inc.

Professor Gregory Abowd
School of Interactive Computing
Georgia Institute of Technology

Dr. Caroline Pantofaru
Google Research
Google Inc.

Professor Thad Starner
School of Interactive Computing
Georgia Institute of Technology

Date Approved: 18 Dec 2015

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor Irfan Essa. This entire body of work would have been impossible without his guidance, patience and the endless hours of work he put into teaching me how to conduct good research and write good papers. Irfan has always been incredibly supportive and he always gave me the freedom to pursue my research interests and explore new collaborations. Looking back, I can confidently say that my years at Georgia Tech have been some of the best years of my life and I have Irfan to thank for it.

I am very grateful to my mentor Caroline Pantofaru who was always there to support and guide me. Her knowledge, energy and passion for solving hard problems were instrumental in shaping my research outlook. She worked tirelessly with me and pushed me to better myself in all aspect of research and development. I thank her for her wonderful mentorship.

I thank my committee members Gregory Abowd, Thad Starner and Rahul Sukthankar, who have generously given their time and expertise to better my work. I thank you for all your contribution and support.

What makes research exciting is the wonderful people we get to collaborate with. I thank senior researchers Thomas Plotz, Grant Schindler, Joshua Jones, Matthias Grundmann and Kihwan Kim and my fellow lab mates and co-authors Edison Thomaz, Daniel Castro, Steve Hickson, Yachna Sharma and Aneeq Zia. I cherish the time we spent together and the wonderful discussions and the fun we had.

A special thanks to my current and former team mates at Google, Arthur Wait, Austin Abrams, Cheol Park, Ezequiel Cura, Julius Kammerl and Krishna Bharat, for their support and encouragement.

Finally, I thank my mom and my family and friends, for all their love and support. Without them, I would not have come this far in life. A very special thanks to the most wonderful person in my life, my wife Shivapriya, who has always supported me through all these years of my studies. Her love, patience and caring helped me get through even the most stressful of times. I am grateful to have her by my side.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS	iii
LIST OF TABLES	viii
LIST OF FIGURES	ix
SUMMARY	xi
I INTRODUCTION	1
1.1 Importance of Context in Scene Understanding	2
1.2 Context In Static Scene Understanding	2
1.3 Context In Dynamic Scene Understanding	3
1.4 Publications and Contributions	5
II LEVERAGING SPATIO-TEMPORAL CONTEXT	9
2.1 Introduction	9
2.2 Related Work	11
2.3 Activity Recognition with Augmented BoW	12
2.3.1 Discovering Temporal Information	13
2.3.2 Encoding Local Structure	15
2.3.3 Capturing Global Structure	17
2.3.4 Activity Recognition	19
2.4 Experimental Evaluation	19
2.4.1 Ocean City Surveillance Data	21
2.4.2 Surgical Skill Assessment	23
2.4.3 Learning Player Activities from Soccer Videos	24
2.4.4 Wide Area Airborne Surveillance (WAAS)	26
2.4.5 Test for Statistical Significance	27
2.4.6 Evaluation Strategy	28
2.5 Conclusion	28

III LEVERAGING EGOCENTRIC CONTEXT	29
3.1 Introduction	29
3.2 Related Work	31
3.3 Egocentric FOV Localization	32
3.3.1 Data collection	32
3.3.2 Reference dataset	33
3.3.3 Matching	33
3.3.4 Correction using sensor data	35
3.3.5 Global Matching and Score Computation	36
3.4 Applications and Results	36
3.4.1 Outdoor Urban Environments	37
3.4.2 Presentations in Indoor Spaces	38
3.4.3 Egocentric Video Tours in Museums	39
3.4.4 Joint Egocentric FOV Localization	41
3.4.5 Evaluation Strategy	42
3.5 Discussion	43
3.6 Conclusion	44
IV LEVERAGING GEOGRAPHIC CONTEXT	46
4.1 Introduction	46
4.2 Related Work	48
4.3 Methodology	49
4.3.1 Image Acquisition	50
4.3.2 Geo-Localizing Images	50
4.3.3 Weakly Supervised Learning	51
4.4 Study & Evaluation	55
4.4.1 Comparative Evaluations	55
4.4.2 Food Recognition in Restaurants	56
4.4.3 Recognition Without Location Prior	59

4.4.4	Evaluation Strategy	60
4.5	Discussion	60
4.6	Conclusion	62
V	LEVERAGING ENVIRONMENTAL CONTEXT	63
5.1	Introduction	63
5.2	Related Work	65
5.3	Methodology	68
5.3.1	Cue 1: Audio	68
5.3.2	OCR: Aligning the Stats With the Broadcast Videos	70
5.3.3	Cue 2: Player Ranking	71
5.3.4	Cue 3: Score Differential	72
5.3.5	Cue 4: Basket Type	73
5.3.6	Cue 5: Motion	74
5.3.7	Cue Combination	75
5.3.8	Generating Highlights	75
5.4	Evaluation	76
5.4.1	Dataset Description	76
5.4.2	Ground-Truth Pairwise Excitement	77
5.4.3	Inter-Rater Reliability	80
5.4.4	Evaluating Individual Cues	80
5.4.5	Evaluating Weighted Cue Combination	82
5.4.6	Evaluating Highlights	84
5.4.7	Evaluation Strategy	88
5.5	Discussion	89
5.6	Conclusion	91
VI	CONCLUSION	92
	REFERENCES	99

LIST OF TABLES

1	Surgical skill assessment using OSATS assessment scheme	24
2	Cluster quality on soccer videos dataset	25
3	McNemar’s tests on statistical significance	27
4	Automated food classification results	57
5	Inter-rater reliability for pairwise excitement A/B tests	80
6	Performance of each individual cue	82
7	Statistical significance of cue combination over each cue	83
8	A/B testing cue combination highlights vs. individual cue highlights with all users	84
9	A/B testing cue combination highlights vs. individual cue highlights with only basketall fans	86

LIST OF FIGURES

1	Histogram of event durations	14
2	Building n -grams and their histograms	15
3	Sample frame from Ocean City data	21
4	Classification results for Ocean City dataset	22
5	Training sessions for surgical skill assessment	23
6	Sample stills from soccer videos dataset	25
7	Results on the WAAS dataset	26
8	An overview of our egocentric FOV localization system	33
9	Egocentric FOV localization in outdoor environments	38
10	Egocentric FOV localization in indoor environments	39
11	Egocentric FOV localization in indoor art installations	40
12	The widths and heights of the 250 paintings	40
13	Joint FOV localization heatmap	43
14	Joint egocentric attention between groups of people	45
15	An overview of our automatic food recognition approach	50
16	Weakly-labeled training images for automated food recognition	51
17	Food segmentation using hierarchical segmentation	52
18	Results on the PFID food dataset	56
19	Testing images for automated food recognition	57
20	Confusion matrices for the 5 different cuisines	58
21	Basketball highlights system overview	68
22	Audio loudness plots for two sample baskets	69
23	Basketball graphics overlay	71
24	Score differential plots for a sample game	72
25	Preference in baskets shown in the highlights based on basket type	73
26	Histogram of duration of the baskets shown in the ESPN clips	76
27	Sample A/B test page shown to the users	78

28	Ground-truth A/B testing: user distribution	79
29	Performance of each individual cue	81
30	Weight of each individual cue in the final cue combination	83
31	Sample frames from highlights generated using individual cues	85
32	Common baskets picks across out highlights and ESPN highlights	87
33	Distribution of the baskets across the two halves of the game	88
34	Summary of the contributions of this thesis	93

SUMMARY

Environments with people are complex, with many activities and events that need to be represented and explained. The goal of scene understanding is to either determine what objects and people are doing in such complex and dynamic environments, or to know the overall happenings, such as the highlights of the scene. The context within which the activities and events unfold provides key insights that cannot be derived by studying the activities and events alone. *In this thesis, we show that this rich contextual information can be successfully leveraged, along with the video data, to support dynamic scene understanding.*

We categorize and study four different types of contextual cues: (1) spatio-temporal context, (2) egocentric context, (3) geographic context, and (4) environmental context, and show that they improve dynamic scene understanding tasks across several different application domains.

We start by presenting data-driven techniques to enrich spatio-temporal context by augmenting Bag-of-Words models with temporal, local and global causality information and show that this improves activity recognition, anomaly detection and scene assessment from videos. Next, we leverage the egocentric context derived from sensor data captured from first-person point-of-view devices to perform field-of-view localization in order to understand the user’s focus of attention. We demonstrate single and multi-user field-of-view localization in both indoor and outdoor environments with applications in augmented reality, event understanding and studying social interactions. Next, we look at how geographic context can be leveraged to make challenging “in-the-wild” object recognition tasks more tractable using the problem

of food recognition in restaurants as a case-study. Finally, we study the environmental context obtained from dynamic scenes such as sporting events, which take place in responsive environments such as stadiums and gymnasiums, and show that it can be successfully used to address the challenging task of automatically generating basketball highlights. We perform comprehensive user-studies on 25 full-length NCAA games and demonstrate the effectiveness of environmental context in producing highlights that are comparable to the highlights produced by ESPN.

CHAPTER I

INTRODUCTION

The human visual system has a remarkable capability of understanding complex and dynamic scenes. Replicating this capability in automated systems has been an active area of research within the computer vision community for the past two decades. A “scene” is the environment where an action or event is occurring and contains active or passive observers. The scene could be categorized as being “dynamic” if the camera and/or the objects and people within the scene are in motion. Examples of dynamic scenes that we use in this thesis include (1) a person wearing an egocentric camera, walking around in a museum gallery and recording point-of-view videos (the camera is in motion while the display pieces in the museum are fixed), (2) a person taking a series of food pictures in a restaurant at regular intervals (the camera position is relatively the same but the object being captured changes shape and size over time) and (3) a camera recording videos of a sports game in a stadium (the camera is fixed while the players in the scene are in motion).

Early research on scene understanding was mostly focused on scene classification, image segmentation and object detection and recognition in static scenes (i.e. scenes in still images). As the state-of-the-art moved towards more challenging real-world “in-the-wild” images and videos, researchers started to look at other cues, such as “context”, that could be leveraged for better scene understanding. Studies of the human visual system have shown that context plays a crucial role in the way we perceive and understand scenes. Let us start with a brief review of the study of context followed by a review of the application of contextual cues for static scene understanding. We then describe context for dynamic scenes, and provide an overview of the

work done in this thesis.

1.1 Importance of Context in Scene Understanding

Contextual cues play a very important role in the human visual system. Humans extensively leverage context by looking at the global organizational structure of the problem rather than making decisions using the immediate local structure [42]. Furthermore, there is strong psychological evidence showing that context plays a crucial role in scene understanding [12]. Useful contextual information can be derived from low level visual features (such as color and texture) sampled over a wide receptive field [128] and this information is used by the human visual system to perform rapid global scene analysis before conducting more detailed local object analysis [84]. Furthermore, this contextual information is available early in the visual processing pipeline and modulates the saliency of image regions and provides an efficient shortcut for object detection and recognition [114].

1.2 Context In Static Scene Understanding

Inspired by the role of context in human visual perception, researchers started studying the context surrounding objects and activities in images. Context in static scenes, as defined in this thesis, is the spatial region surrounding the region-of-interest (ROI), which could be a larger image patch around the ROI or the entire scene contained within the image. In a given image, context is provided by the type of the scene, the objects within the ROI and their relationship with each other. This can be studied by jointly modeling and reasoning about objects and scenes [117, 107] and building spatial context-aware models [103]. Local patch-based object detection can be improved and local ambiguities can be resolved by looking at the whole image to get the global context of the scene [83]. The relationship between objects and other regions in the image can be modeled using Conditional Random Fields [60] and object detection can be improved by estimating the rough 3D scene geometry from a single

image [49]. Contextual information can also be used to go beyond object detection and hierarchical models can be built to recognize sporting events in static images [68].

Apart from the contextual information provided by the relationship between objects and scenes, the presence of other objects and the relationship between the objects within a scene also provides valuable contextual cues. Previous work has studied the contextual information provided by the interdependence of objects, surface orientations and camera viewpoints [50], geometric consistency of objects [31], geometric relationships between objects at local and global scales [45], joint spatial constraints between objects in 2D and 3D imagery [106] and by modeling the relationship between objects and other objects using Discriminative Random Fields (DRFs) [59]. Moving beyond geometric and spatial relationships, semantic context has also been explored by using “prepositions” and “comparative adjectives” instead of just “nouns” to express the relationship between objects [35], using contextual constraints such as co-occurrence of objects in sports scenes for tasks such as classification, annotation and segmentation [69] and by looking at the semantic relations between objects as post-processing to off-the-shelf object categorization approaches [94]. Another interesting work studied mutual boosting with multiple detectors of objects and parts (trained in parallel using AdaBoost) for incorporating contextual information to improve object detection [115].

Supported by the above listed advances in the state-of-the-art for static scene understanding, we next explore, categorize and leverage contextual cues for dynamic scenes in videos.

1.3 Context In Dynamic Scene Understanding

While the term “context” has multiple meanings in scene classification literature, for dynamic scenes it almost always refers to the “spatio-temporal context”. This is the information contained within the space-time video volume that encapsulates low-level

visual features that are computed using local and semi-local statistics. Research in activity recognition from videos has explored spatio-temporal context in the form of Bag-of-Words (BoW) models [126] and more recently with robust descriptors, which exploit continuous object motion and integrate it with distinctive appearance features [21], spatio-temporal features based on dense trajectories [125] and features learnt in an unsupervised manner directly from video data [64].

Our research is aimed at expanding “spatio-temporal” context for videos by building more informative BoW models and exploring other types of contextual cues such as “egocentric context”, “geographic context” and “environmental context” for better dynamic scene understanding. These different types of context can be broadly categorized into three classes:

1. Context that is part of the video data: This is the “spatio-temporal” context that is contained within the space-time video volume that surrounds the visual words of an activity. This type of context is baked into the video data and has to be explicitly extracted and modeled. The vast majority of research in leveraging context for dynamic scene understanding has focused on this particular context class.

2. Context that is concurrently collected from an external device co-located with the camera: This is usually the contextual information obtained from various sensors located on (or in close proximity to) the camera capturing the activity. Examples include “egocentric” context obtained from a first-person point-of-view devices such as Google Glass that describes the person’s head orientation information using devices such as accelerometers, gyroscopes and magnetometers; and “geographic context”, obtained from GPS data that describes the location where the activity is taking place. This type of context is captured concurrently with the video data and can be treated as an additional stream of information.

3. Context provided by the environment: This is the “environmental” context, captured using sensors in the environment where the activity is taking place,

such as audio, video and additional meta-data obtained from third-party observers in the scene. The environmental context changes over time as factors in the environment react to the ongoing activity.

Given these types of context, the thesis statement can be formulated as follows: *Contextual information can be extracted from the data, collected from external sensors or gathered from the environment and can be effectively leveraged, along with the temporally varying data, to improve dynamic scene understanding.* The following chapters are devoted to the development, support and explanation of this thesis statement with supporting evidence from several different application domains.

1.4 Publications and Contributions

The following publications form the basis for this thesis:

- “Augmenting Bag-of-Words: Data-Driven Discovery of Temporal and Structural Information for Activity Recognition”, Vinay Bettadapura, Grant Schindler, Thomas Ploetz, Irfan Essa, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013*
- “Egocentric Field-of-View Localization Using First-Person Point-of-View Devices”, Vinay Bettadapura, Irfan Essa, Caroline Pantofaru, *IEEE Winter Conference on Applications of Computer Vision (WACV), 2015 (won the best paper award)*
- “Leveraging Context to Support Automated Food Recognition in Restaurants”, Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory Abowd, Irfan Essa, *IEEE Winter Conference on Applications of Computer Vision (WACV), 2015*

- “Automatic Generation of Basketball Highlights Using Contextual Cues”, Vinay Bettadapura, Caroline Pantofaru, Irfan Essa, *In Submission*

Our contributions are as summarized below:

Leveraging Spatio-Temporal Context: While state of-the-art BoW models are good at building powerful representations of the data, they completely ignore the ordering of the particular words regarding their absolute and relative positions. Furthermore, standard BoW approaches do not account for the fact that different types of activities have different temporal signatures. We describe a method to represent temporal information by quantizing time and defining new temporal events in a data-driven manner. We propose three encoding schemes that use n -grams to augment BoW with the discovered temporal events in a way that preserves the local word-ordering information (relative word positions) in the activity. In addition, to discover the global contextual patterns in the data, we introduce randomly sampled Regular Expressions to augment our BoW models. Our proposed approach is evaluated on four challenging scene-understanding tasks: vehicle activity recognition, surgical skill assessment, unsupervised learning of player roles in soccer games and recognition of human behaviors and anomaly detection in massive surveillance datasets.

Leveraging Egocentric Context: A key question in dynamic scene understanding: “What is the user looking at right now?” Automatically analyzing the point-of-view (POV) data (images, videos and sensor data) to estimate egocentric perspectives and shifts in the FOV remains challenging. Due to the unconstrained nature of the data, no general FOV localization approach is applicable for all outdoor and indoor environments. Our insight is to make such localization tractable by introducing a reference dataset (i.e., a visual model of the environment, which is pre-built or concurrently captured, annotated and stored permanently) and to leverage the egocentric context (which is the captured POV data from one or more egocentric devices) by matching and correlating it against this reference data-set allowing for transfer of

information from the user’s reference frame to a global reference frame of the environment. The problem is now reduced from an open-ended data-analysis problem to a more practical data-matching problem. Our method is applied to outdoor urban environments, indoor presentations, egocentric tours in museums and analyzing joint egocentric attention of groups of people.

Leveraging Geographic Context: We look at the challenging problem of “in-the-wild” food recognition and show that leveraging geographic context helps make the problem tractable [11]. We present an automatic workflow where online resources are queried with contextual sensor data such as GPS to find food images and additional information about the restaurant where the food picture was taken, with the intent to build classifiers for food recognition. Our method is evaluated on food images taken in 10 restaurants across 5 different types of cuisines (American, Indian, Italian, Mexican and Thai). Using food and restaurants as the domain, we demonstrate the value of geographic context for dynamic scene understanding. We believe that the same method can be used for many other domains.

Leveraging Environmental Context: Sporting events such as basketball are classic examples of dynamic scenes. The players are active and constantly on the move, the audience is excited and cheering and there are several third party observers within the scene such as referees, coaches, commentators and on-court statisticians. Sporting environments such as stadiums and gymnasiums have broadcast cameras that provide us with the video feed and are a rich source of environmental contextual cues. The environment (the audience and the third-party observers within the scene) reacts to the game as the game progresses. The audience cheer and the meta-data provided by the on-court statisticians provide valuable information that can be leveraged to understand the sporting scene. An effective demonstration of dynamic scene understanding is the automated production of basketball highlights. We analyzed 25

full-length NCAA games and explored the use of four different environmental contextual cues (“Score Differential”, “Player Ranking”, “Basket Type” and “Commentator and Audience Audio”) along with a fifth cue derived from the video (“Motion”). Our extensive user studies conducted using Mechanical Turk show the effectiveness of combining these five cues in producing highlights that are comparable to the ones produced by ESPN.

The following chapters describes each of the above contributions in detail.

CHAPTER II

LEVERAGING SPATIO-TEMPORAL CONTEXT

We present data-driven techniques to enrich the spatio-temporal context provided by Bag of Words models (BoW) by introducing Augmented Bag of Words (ABoW), which allow for more robust modeling and recognition of complex long-term activities, especially when the structure and topology of the activities are not known a priori [10]. Our approach specifically addresses the limitations of standard BoW approaches, which fail to represent the underlying temporal and causal information that is inherent in activity streams. In addition, we also propose the use of randomly sampled regular expressions to discover and encode patterns in activities. We demonstrate the effectiveness of our approach in experimental evaluations where we successfully recognize activities and detect anomalies in four complex datasets.

2.1 Introduction

Activity recognition in large, complex datasets has become an increasingly important problem. Extracting activity information from time-varying data has applications in dynamic scene understanding such as video understanding, scene assessment and surveillance for anomaly detection. Traditionally, sequential models like Hidden Markov Models (HMMs) and Dynamic Bayesian Networks have been used to address activity recognition as a time-series analysis problem. However, the assumption of Markovian dynamics restricts the application of such sequential models to relatively simple problems with known spatial and temporal structure of the data to be analyzed [118]. Similarly, syntactic methods like Parse Trees and Stochastic Context Free Grammars [82, 54] are not well suited for recognizing weakly structured activities and are not robust to erroneous or uncertain data.

As a promising alternative, research in activity recognition from videos and other time-series data started incorporating spatio-temporal contextual information by moving towards bag-of-words (BoW) approaches and away from the traditional sequential and syntactic models. However, while BoW approaches are good at building powerful and sparser representations of the data, they completely ignore the ordering and structural information of the particular words regarding their absolute and relative positions. Furthermore, standard BoW approaches do not account for the fact that different types of activities have different temporal signatures. Each event in a long-term activity has a temporal duration, and the time that passes between each pair of consecutive events, is different for different activities

In this thesis, we expand the role of spatio-temporal context by introducing novel BoW techniques and extensions that explicitly encode the temporal and structural information gathered from the data. Recent activity recognition approaches such as [86] have extended the BoW approach with topic models [122] using probabilistic Latent Semantic Analysis [48] and Latent Dirichlet Allocation [13], leading to more complex classification methods built on top of standard BoW representations. In contrast, we increase the richness of the spatio-temporal context in the BoW representation and with the use of standard classification backends (like k -NN, HMM and SVM), we demonstrate that our augmented BoW techniques lead to better recognition of complex activities.

Contributions: We describe a method to represent temporal information by quantizing time and defining new temporal events in a data-driven manner. We propose three encoding schemes that use n -grams to augment BoW with the discovered temporal events in a way that preserves the local structural information (relative word positions) in the activity. This narrows the conceptual gap between BoW and sequential models. In addition, to include more spatio-temporal context through the

discovery the global patterns in the data, we augment our BoW models with randomly sampled Regular Expressions. This sampling strategy is motivated by the random subspace method as it is commonly used for decision tree construction [14] and related approaches which have shown success in a wide variety of classification and visual recognition problems [67].

We evaluate our approach in comparison to standard BoW representations on four diverse classification tasks: *i*) Vehicle activity recognition from surveillance videos (Section 2.4.1); *ii*) Surgical skill assessment from surgery videos (Section 2.4.2); *iii*) Unsupervised learning of player roles in soccer videos (Section 2.4.3) and *iv*) Recognition of human behavior and anomaly detection in massive wide-area airborne surveillance (simulation) data (Section 2.4.4). Recognition using our augmented BoW outperforms the standard BoW approaches in all four datasets. We provide evidence that this superior performance generalizes to any classification framework by demonstrating how sequential models (HMMs), instance based learning (k -NNs), and discriminative recognition techniques (SVMs) benefit from the new representation and outperform respective models trained on standard BoW. Finally, we show how augmented BoW-based techniques successfully unveil further details of the analyzed datasets, such as behavior anomalies.

2.2 Related Work

The Bag of Words (BoW) model was first introduced for Information Retrieval (IR) with text [96]. Since then, it has been used extensively for text analysis, indexing and retrieval [75]. Building on the success of BoW approaches for IR with text and images, research in activity recognition has focused on working with BoW built using local spatio-temporal features [124] and more recently with robust descriptors, which exploit continuous object motion and integrate it with distinctive appearance features [20], features based on dense trajectories [123] and features learnt in an unsupervised

manner directly from video data [63].

While the focus has mostly been on recognizing human activities in controlled settings, recent BoW based approaches have focused on recognizing human activities in more realistic and diverse settings [62], and with the use of higher level semantic concepts (attributes) that allow for more descriptive models of human activities [72]. However, when activities are represented as bags of words, the underlying sequential information provided by the ordering of the words is typically lost. To address this problem, n -grams have been used to retain some of the ordering by forming sub-sequences of n items [75] (Figure 2). More recently, variants of the n -gram approach have been used to represent activities in terms of their local event sub-sequences [37]. While this preserves local sequential information and causal ordering, adding absolute and relative temporal information results in more powerful representations as we demonstrate in this chapter.

Our augmentation method is independent of the underlying BoW representation, i.e., the modality of the data to be processed. The input to our algorithm is a sequence of atomic events, i.e., words. On video data these can be either derived from state-of-the-art short-duration event detectors (e.g., the Actom Sequence Model [32], automatic action annotation [23]), or any other suitable feature detectors.

2.3 Activity Recognition with Augmented BoW

We define an *activity* as a finite sequence of events over a finite period of time where each *event* in the activity is an occurrence. For example, if “start”, “turn”, “straight” and “stop” are four individual events, then a vehicle driving activity will be a finite sequence of those events over some finite time (e.g. “start → straight → turn → stop → start → straight → stop”). We call these events, that can be described by an observer and have a semantic interpretation, as *observable events*.

Recent methods for activity recognition try to detect such observable events and

build BoW upon it. However, the temporal structure underlying the activities that shall be recognized is typically neglected. The time taken by each observable event and the time elapsed between two subsequent events are two important properties that contribute to the temporal signature of an activity that is being performed. For example, a car at a traffic light will have a shorter time gap between the “stop” and “start” events than a delivery vehicle that has to stop for a much longer time (until its contents are loaded/unloaded) before it can start again.

2.3.1 Discovering Temporal Information

We represent activities as sequences of discrete, observable events. Let $\omega = \{a_1, a_2, a_3, \dots, a_p\}$ denote a set of p activities, and let $\phi = \{e_1, e_2, e_3, \dots, e_q\}$ denote the set of q types of observable events. Each activity a_i is a sequence of elements from ϕ . Each event type can occur multiple times at different positions in a_i .

We now introduce *temporal events*. Let $\tau_{j,k}$ be the temporal event defined as the time elapsed between the end of observable event e_j and the start of observable event e_k , where $k > j$. Since it measures time, $\tau_{j,k}$ is non-negative. Also, let $\pi_{j,k}$ be the temporal event defined as the time elapsed between the start of observable event e_j and the end of observable event e_k , where $k \geq j$. Thus, $\tau_{j,k}$ measures the time elapsed between any two events whereas $\pi_{j,k}$ measure the time elapsed between any two events including the time taken by those two events. Thus, $\tau_{j,k}$ and $\pi_{j,k}$ are related by the equation $\pi_{j,k} = \pi_{j,j} + \tau_{j,k} + \pi_{k,k}$. We posit that these two types of temporal events, $\tau_{j,k}$ and $\pi_{j,k}$, can model all the temporal properties of an activity. The four possible scenarios are listed here:

1. $\tau_{j,j+1}$: Time elapsed between any two consecutive events e_j and e_{j+1}
2. $\tau_{j,k}$: Time elapsed between any two events e_j and e_k , where $k > j$
3. $\pi_{j,j}$: Time taken by a single event e_j

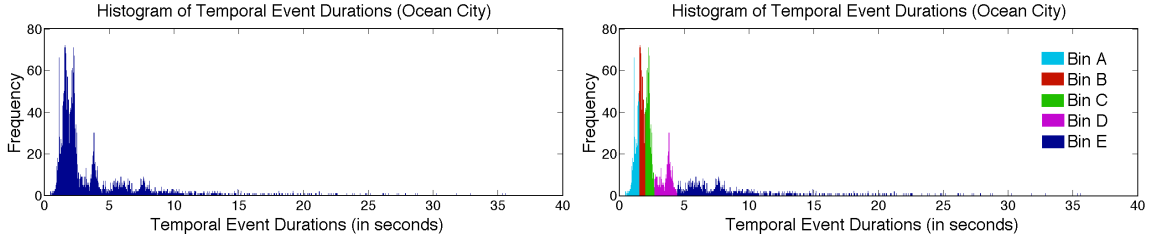


Figure 1: Histogram of event durations for Ocean City dataset (left) and data-driven creation of temporal bins (right; $N = 5$).

4. $\pi_{j,k}$: Time taken by set of events e_j to e_k , where $k \geq j$

To work with these temporal events, we will have to quantize them into a finite number of N bins. This quantization is crucial in allowing us to incorporate a notion of time into BoW models. However, uniformly dividing the time-line into N bins is not ideal. As illustrated by the temporal event duration histograms of $\tau_{j,j+1}$ for the Ocean-City dataset (see Section 2.4.1) in Figure 1, short and medium duration temporal events occur much more frequently than longer duration temporal events. Similar temporal distributions are observed in the other datasets we have analyzed.

To ensure that we capture the most useful temporal information, we pursue a data-driven approach for binning. Bins are selected based on the distribution of temporal events. If there are S temporal events, then we divide the temporal space into N bins such that each of the N bins contains an equal proportion S/N of the temporal events (illustrated in Figure 1 for $N = 5$). Note that, if the time-line had been naively divided into 5 equally sized bins, then most of the temporal events would have been placed in the first bin while the other 4 bins would have been almost empty. The choice of N depends on the problem we are addressing. Lower values of N result in increased loss of temporal information.

Example 1: Say, temporal event $\tau_{j,k}$ is of 4 second duration and temporal event $\tau_{l,m}$ is of 20 second duration, then from Figure 1, we see that $\tau_{j,k}$ will be assigned to bin D and $\tau_{l,m}$ will be assigned to bin E . Let ψ denote the function that maps the temporal events to their respective temporal bins. Then, we can say that $\psi(\tau_{j,k}) = D$

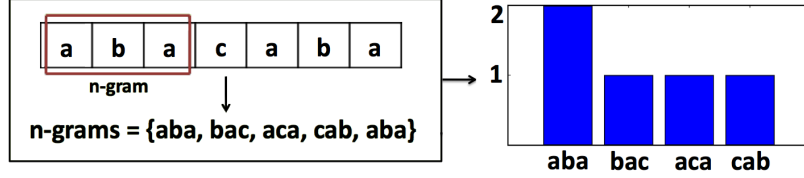


Figure 2: Building n -grams and their histogram (here $n = 3$) [37]

and $\psi(\tau_{l,m}) = E$.

There are many possible ways by which we can encode these new temporal events along with the observable events to build augmented BoW representations. The simplest way would be to just add the quantized temporal events to the BoW, i.e., if the BoW contained x observable events and we extracted y new quantized temporal events, then the augmented BoW will now contain $x+y$ number of elements. Although this naive representation already gives better results than just the BoW (see Section 2.4), as shown in the next section, more sophisticated alternatives are possible.

2.3.2 Encoding Local Structure

In the following we describe three encoding schemes we have developed that merge the temporal events with the observable events in a way that captures local structure.

2.3.2.1 Interspersed Encoding

In interspersed encoding, the main focus is on the time elapsed between every pair of consecutive events. Let $\tau_{j,j+1}$ be a temporal event defined as the time elapsed between any two consecutive observable events e_j and e_{j+1} in activity a_i . Once the quantized temporal events $\psi(\tau_{j,j+1})$ are computed for all event pairs $e_j, e_{j+1} \in a_i$, they are then inserted into a_i at their appropriate positions between events e_j and e_{j+1} . Let this new sequence of *interspersed* events for activity a_i be denoted by T_i . In general, if activity a_i has d events, then after the inclusion of the quantized temporal events, T_i will have $2d - 1$ events (the original d observable events plus the new $d - 1$ temporal events).

Example 2: For the activity $a_1 = (e_1, e_2, e_3)$, we have $T_1 = (e_1, \psi(\tau_{1,2}), e_2, \psi(\tau_{2,3}), e_3)$. If temporal event $\tau_{1,2}$ is of 4 second duration and $\tau_{2,3}$ is of 20 second duration, then the quantized temporal events will be $\psi(\tau_{1,2}) = D$ and $\psi(\tau_{2,3}) = E$. So, the interspersed sequence of events for activity a_1 , will be $T_{1=} = (e_1, D, e_2, E, e_3)$.

One of the main drawbacks of classical BoW representations is the loss of original word orderings (i.e. local structural information). This is particularly adverse in the context of activity recognition because activities correspond to causal chains of observable and temporal events. Losing the ordering will result in a loss of all causality and contextual information. We employ n -grams in order to retain ordering of events [28]. An n -gram is a sub-sequence of n terms from a given sequence. Deriving n -grams and their histograms from a given sequence is illustrated in Figure 2.

Using this approach, for every activity a_i , the event sequence T_i is transformed into an n -gram sequence T_i^I (where the superscript I stands for *interspersed*). This T_i^I feature vector representing activity a_i is the final result of interspersed encoding. From Example 2, with $n = 3$, the event sequence $T_{1=} = (e_1, D, e_2, E, e_3)$ will be transformed into the n -gram sequence $T_1^I = (e_1De_2, De_2E, e_2Ee_3)$ or in its histogram form $T_1^I = \{e_1De_2 \Rightarrow 1, De_2E \Rightarrow 1, e_2Ee_3 \Rightarrow 1\}$ (denoted as key-value pairs where the key is the n -gram and the value is its frequency).

2.3.2.2 Cumulative Encoding

In cumulative encoding, the main focus is on the cumulative time taken by a subsequence of observable events. Let $\psi(\pi_{j,j+n-1})$ be a quantized temporal event defined as the total time taken by n consecutive events e_j to e_{j+n-1} in activity a_i . Once the quantized temporal event $\psi(\pi_{j,j+n-1})$ is computed for the consecutive sequence of observable events $e_j \dots e_{j+n-1} \in a_i$, it is appended to the set of the observable events. Let this new sequence of “cumulative” observable and temporal events for activity a_i be denoted by T_i^C (where the superscript C stands for “cumulative”).

Example 3: If activity $a_2 = (e_1, \dots, e_5)$, $n = 3$, then $T_2^C = (e_1e_2e_3\psi(\pi_{1,3}), e_2e_3e_4\psi(\pi_{2,4}), e_3e_4e_5\psi(\pi_{3,5}))$. Say, $\pi_{1,3}$ is of 4 second duration, $\pi_{2,4}$ is of 20 second duration and $\pi_{3,5}$ is of 1 second duration and that $\psi(\pi_{1,3}) = D$, $\psi(\pi_{2,4}) = E$ and $\psi(\pi_{3,5}) = A$. So, the new sequence of events for activity a_2 , will be $T_2^C = (e_1e_2e_3D, e_2e_3e_4E, e_3e_4e_5A)$ or in histogram form $T_2^C = \{e_1e_2e_3D \Rightarrow 1, e_2e_3e_4E \Rightarrow 1, e_3e_4e_5A \Rightarrow 1\}$.

Interspersed encoding focuses on the time elapsed between events whereas cumulative encoding focuses on the time taken by the events.

2.3.2.3 Pyramid Encoding

Given the choice of encoding scheme—either interspersed or cumulative—in pyramid encoding all l -grams of length $l, \forall l \in [1, n]$ are generated. Then we build a pyramid of these l -grams allowing for processing of event sequences at multiple scales of resolution. We denote BoW representations for activity a_i generated through pyramid encoding by T_i^P .

The output of each of these encoding schemes, i.e., T_i^I , T_i^C and T_i^P is the augmented BoW model containing the observable and temporal events, encoded in a way that captures the local structure.

2.3.3 Capturing Global Structure

While n -grams are good at capturing local information, their capability to capture longer range relationships are rather limited. This is where regular expressions come into play. Obviously, it is computationally intractable to enumerate all possible regular expressions for a given vocabulary of observable and temporal events. Thus, given the set of observable events ϕ and the set of discovered temporal events N , we construct a vocabulary of all events $\phi \cup N$ denoted by Γ where $|\Gamma| = |\phi| + |N|$, and create a sub-space of regular expressions by restricting their form to:

$$\wedge . * (\alpha) (\beta_1 | \dots | \beta_r) \varphi (\gamma) . * \$ \quad (1)$$

where the symbols $\alpha, \beta_i, \gamma \in \Gamma$ with $i \in [1, r]$ and $r = \text{rand}(1, |\Gamma|)$. The symbol φ is randomly set to one of the three quantifier characters: $\{*, +, ?\}$. The special characters have the following meaning: “ \wedge ” matches the start of the sequence, “ $.$ ” matches any element in the sequence, “ $*$ ” matches the preceding element zero or more times, “ $+$ ” matches the preceding element one or more times, “ $?$ ” matches the preceding element zero or one time and “ $\$$ ” matches the end of the sequence. The “ $|$ ” operator matches either of its arguments. For example, $e_1(e_2|e_3)e_4$ will match either $e_1e_2e_4$ or $e_1e_3e_4$.

The first symbol that will be matched (α) and last symbol that will be matched (γ) are chosen randomly from Γ using probability-proportional-to-size sampling (PPS) and the r intermediate symbols β_i are chosen randomly from Γ using simple random sampling (SRS). PPS concentrates on frequently occurring events and picks the first and last symbols in the regular expression to be the ones that have the greatest impact on the population estimates whereas SRS chooses each of the intermediate symbols with equal probability, thus giving a fair chance for all events to equally participate in the matching process. The results of our experimental evaluation suggest that this combination of PPS-SRS sampling of the regular expression subspace strikes the right balance between discovering global patterns across activities and discovering the anomalous activities.

Regular expressions of the above form are randomly generated and those that do not match at least one of the activities/event-sequences are rejected. Accepted regular expressions are treated as new words and added to our augmented BoW representation. This final representation now contains automatically discovered temporal information and both local and global structural information of the activities. Our

experiments show that increasing the number of words in BoW through randomly generated regular expressions by just 20% boosts the activity recognition and anomaly detection results significantly (Section 2.4).

2.3.4 Activity Recognition

Activity recognition using augmented BoW is pursued in a straightforward manner by feeding the time-series data in their novel representation into statistical modeling backends. Note that there is in principle no limitation on the kind of classification framework to be employed. In Section 2.4 we present results for instance based learning (k -NN), sequential modeling (HMM), and discriminative modeling (SVM).

Given videos or time-series data of activities, temporal information is discovered using the histogram method described in Section 2.3.1. Using n -grams, the temporal information is then merged with the extracted BoW thereby preserving local ordering of the words. The new BoW model is then further augmented by adding new words created using randomly sampled regular expressions (to capture global patterns in the data), and then processed by the statistical modeling backend for actual activity recognition.

2.4 *Experimental Evaluation*

The methods presented in this chapter were developed in order to improve BoW-based activity recognition, thereby aiming for generalization across application domains. For practical validation, we have thus evaluated our approaches in a range of experiments that cover three diverse classes of learning problems (binary classification, multi-class classification, and unsupervised learning) across four challenging datasets from different domains.

Optimization of the estimation procedure for augmented BoW representations involves the two main parameters in our system: N , the number of temporal bins used for quantization and n , the size of the n -gram used for encoding. Low values

of N and n result in the loss of temporal and structural information whereas high values can lead to large BoW with very high dimensionality. The optimal values for N and n are determined by standard grid search [18]. Within a user supplied interval, all grid points of (N, n) are tested to find the combination that gives the highest accuracy. 50% of the particular datasets is held-out for parameter optimization, and the remaining 50% is used for model estimation using cross-validation. This provides an unbiased estimate of the generalization error and prevents over-fitting.

The main evaluation criterion for all activity recognition experiments is classification accuracy, which we report as absolute percentages and, for more detailed analysis, in confusion matrices. For the first set of experiments (Section 2.4.1) we compare three different classification backends (k -NNs with cosine-similarity distance metric, HMMs, and SVMs) and explore their capabilities in systematic evaluations of their parameter spaces. Due to space constraints the presentation of results for the remaining set of experiments is limited to those achieved with the k -NN classification backend. These results are, however, representative for all three types of classifiers evaluated.

k -NNs with cosine-similarity distance metric, i.e. Vector Space Models (VSM), treat the derived BoW vectors of activities as document vectors and allow for automatic analysis in terms of querying, classifying, and clustering the activities [75]. Prior to classification, each term in our augmented BoW is assigned a weight based on its term-frequency and document-frequency in order to obtain a statistical measure of its importance. Classification is done using leave-one-out cross-validation (LOOCV).

HMM-based experiments employ semi-continuous modeling with Gaussian mixture models (GMM) as feature space representations [28]. GMMs are derived by means of an unsupervised density learning procedure. All HMMs are based on linear left-right topologies with automatically derived model lengths (based on training data statistics), and are trained using classical Baum-Welch training. Classification is

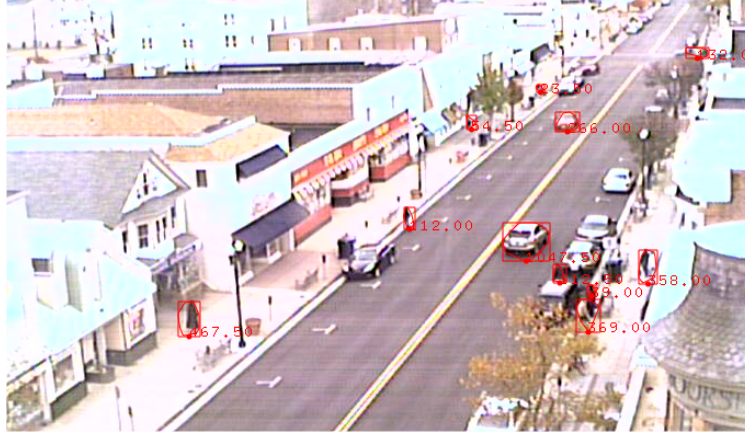


Figure 3: Sample frame from Ocean City data showing the various objects being tracked.

pursued using Viterbi-decoding. Parameter estimation and model evaluation employs 10-fold cross-validation.

Experiments on SVMs are carried out in 10-fold cross-validation using LIBSVM with an RBF kernel. Parameter optimization utilizes a grid-search procedure as it is standard for finding optimal values for C and γ [18].

2.4.1 Ocean City Surveillance Data

The first dataset consists of 7 days of uncontrolled videos recorded at Ocean City, USA [89]. The input video was stabilized and geo-registered and 2,140 vehicle tracks were extracted using background subtraction and multi-object tracking [89] (Figure 3). An event detector analyzed the tracks, detected changes in structure over time and represented each track by a sequence of observable events. The types of events detected in each track were “start”, “stop”, “turn” and “u-turn”.

Out of the 2,140 vehicle tracks, 448 vehicles are either entering or exiting parking areas on either side of the road (Figure 3). The recognition objective is to determine whether or not vehicles are involved in parking activities.

With the empirically determined optimal values of $N = 2$ and $n = 2$, we perform binary classification. The results are shown in Figure 4. For k -NN based experiments,

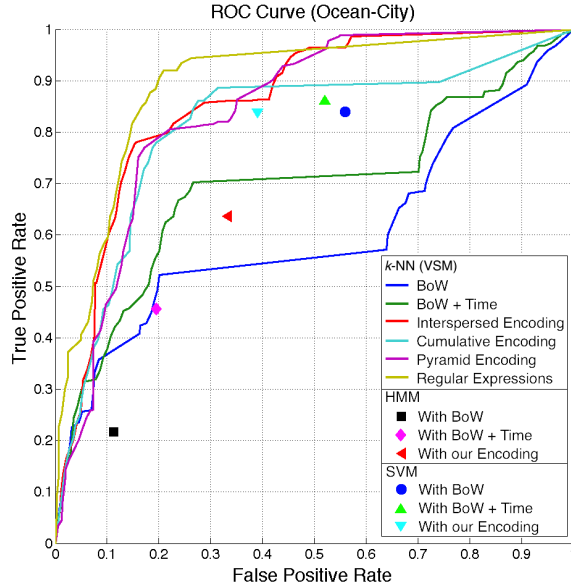


Figure 4: Classification results for Ocean City dataset. Our encoding schemes outperform the BoW baseline on three classification backends: VSM, sequential models (HMMs) and SVMs.

ROC curves were generated by varying the acceptance threshold. Augmenting BoW with temporal information (bag-of-words + time) improves the results over the BoW baseline. The performance is further improved with our proposed Interspersed, Cumulative and Pyramid encoding schemes. However, the best results are obtained when we augment our BoW with randomly generated regular expressions.

Figure 4 also shows the performance of HMM and SVM based recognition backends using augmented BoW representations. Both techniques produce fixed decisions based on maximizing models' posterior probabilities, i.e., no threshold-based post-processing is applied for the actual recognition. Consequently, ROC curves are not applicable, and the particular results are shown as points in the figure.

Analyzing the evaluation results, it becomes evident that: *i*) our proposed encoding schemes outperform the BoW baseline; and *ii*) superior classification accuracy generalizes across recognition approaches (k -NN, HMM, SVM), with largest gain for Vector Space Models.

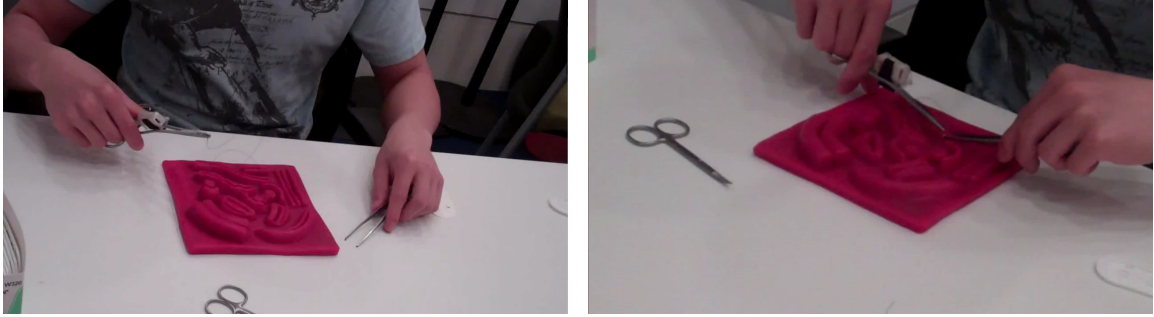


Figure 5: Long-range (left) and close-up (right) stills of video footage from training sessions for surgical skill assessment. Participants practice suturing using regular instruments and suture pads.

2.4.2 Surgical Skill Assessment

The second set of experiments is related to evaluating surgical skills as it is standard routine in practical training of medical students. As part of a larger case-study, 16 medical students were recruited to perform typical suturing activities (stitching, knot tying, etc.) using regular instruments and tissue suture pads. Both long-range and close-up videos of these “suturing” procedures were captured at 50 fps at a resolution of 720p (sample still images in Figure 5). As part of the training procedure, participants completed 2 sessions with 2 attempts in each session, resulting in a total of 64 videos. Ground truth annotation was done by an expert surgeon who assessed the skills of the participants using a standardized assessment scheme (OSATS [76]) based on 7 different metrics (Table 1) on a three-point scale (low competence, medium, and high skill).

Harris3D detectors and histogram of optical-flow (HOF) descriptors [124] are used to extract visual-words from the surgery videos. BoW are built with vocabularies constructed using k -means clustering (with $k = 50$), and then augmented using our techniques. Table 1 summarizes our experiments (using k -NN classification backend) and gives comparisons with the BoW baseline. It can be seen that augmented BoW based approaches outperform the BoW baseline in all 7 skill metrics with an overall accuracy of 72.56%.

Table 1: Surgical skill assessment using OSATS assessment scheme [76]. Ground truth annotation provided by an expert surgeon who assessed the training sessions using 7 different metrics (rows) and a three-point scale (low competence, medium, and high skill). Results given are accuracies from automatic recognition using k -NN, replicating expert assessment based on video footage of the training sessions. Our encoding (Interspersed encoding with 3-grams, 5 time bins and with 20 random regular expressions) outperforms the BoW baseline on all 7 metrics.

	M1: BOW baseline	M2: BOW + Time	M3: Our encoding
Respect for tissue	66.67%	69.84%	73.02%
Time and motion	50.79%	66.67%	74.60%
Instrument handling	50.79%	65.08%	68.25%
Suture handling	69.84%	69.84%	73.02%
Flow of operation	49.21%	63.49%	66.67%
Knowledge of procedure	60.32%	74.60%	80.95%
Overall performance	52.38%	68.25%	71.43%
Average accuracy	57.14%	68.25%	72.56%

Since our augmented BoW representations capture time and co-occurrence of words, we hypothesized that an automated analysis procedure using augmented BoW should perform particularly well in assessing the “time and motion” and “knowledge of procedure” skills. Recognition results reported in Table 1 indicate that this is indeed the case. The classification accuracies are 74.60% and 80.95% (an increase of 23.81% and 20.63% respectively, over the BoW baseline), thus validating our hypothesis.

2.4.3 Learning Player Activities from Soccer Videos

Automatic detection, tracking and labeling of the players in soccer videos is critical for analyzing team tactics and player activities. Previous work in this area has mostly focussed on detecting and tracking the players, recognizing the team of the players using appearance models and detecting short-duration player actions. In our experiments, we consider the problem of unsupervised learning of long-range activities and roles the various players take on the field. Given their tracks, we cluster them into 7 clusters: “Team-A-Goalkeeper”, “Team-A-Striker”, “Team-A-Defense”, “Team-B-Goalkeeper”, “Team-B-Striker”, “Team-B-Defense” and “Referee”.

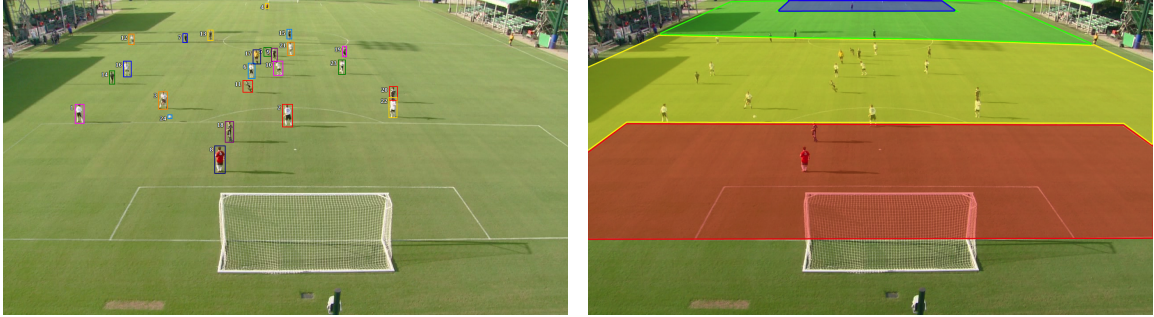


Figure 6: Sample stills from soccer videos dataset. **Left:** The 24 objects being tracked: 22 players from both teams, referee and the ball. **Right:** The 4 zones used by our event detector: Zone-A (Red), Zone-B (Yellow), Zone-C (Green) and Zone-D (Blue).

We analyzed full length match videos (720p at 59.94 fps) from the Disney Research soccer games dataset and tracked the 24 objects (players, referee, and ball) on the field using a multi-agent particle filter based framework [36] (Figure 6). The tracks were given to an event detector that divided the field into 4 zones (Figure 6) and detected 10 types of events: “Enter-Zone-A”, “Leave-Zone-A”, “Enter-Zone-B”, “Leave-Zone-B”, “Enter-Zone-C”, “Leave-Zone-C”, “Enter-Zone-D”, “Leave-Zone-D”, “Receive-Ball” and “Send-Ball”. With this vocabulary of 10 events, we built augmented BoW and clustered them using k -means clustering where $k = 7$. Clustering results are given in Table 2. It can be seen that augmented BoW outperform the BoW baseline on all 3 cluster quality metrics. In a supervised setting, we achieve an accuracy of 82.61%, which is a 17.39% improvement over the BoW baseline (which is 65.22%).

Table 2: Cluster quality on soccer videos dataset. The 3 metrics used are Rand Index (RI), Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Our encoding (Interspersed encoding with 3-grams, 3 time bins and 20 random regular expressions) gives better cluster quality than the BoW baseline.

	RI	ARI	NMI
BOW baseline	0.7984	0.2922	0.6147
BOW + Time	0.8300	0.3920	0.6974
Our Encoding	0.8261	0.5244	0.7462

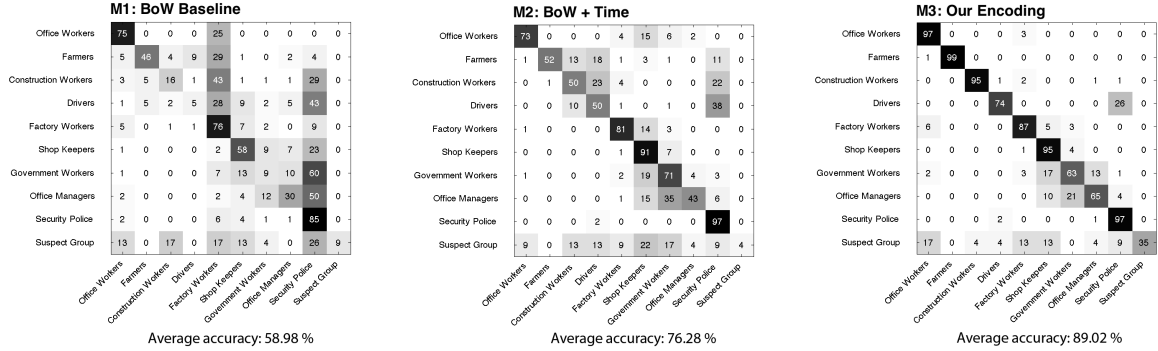


Figure 7: Results on WAAS dataset: **Left:** BoW baseline; **Middle:** BoW + Time; **Right:** Our encoding (5-grams, 5 time bins and with 1,000 random regular expressions). Overall improvement of 30.04% is observed with our method (compared to standard BoW baseline).

2.4.4 Wide Area Airborne Surveillance (WAAS)

In order to evaluate the applicability and scalability of our approach on massive datasets with several hundreds of thousands of activities, we consider the Wide-Area Airborne Surveillance (WAAS) simulation dataset.

The WAAS dataset was developed by the U.S. Military as part of their Activity Based Intelligence (ABI) initiative. The goal is to capture motion imagery from an airborne platform that provides persistent coverage of a wide area, such as a town or a small city, and merge the automatically captured data from the aerial station with intelligence gathered by ground forces to build a surveillance database of humans and vehicles in that area. In order to aid research in this area, the WAAS dataset has been released, which contains Monte Carlo simulation of the activities of 4,623 individuals for a total duration of 46.5 hours generated in 1 minute increments. There are a total of 180 events (like “Eat Lunch”, “Enter Vehicle”, “Exit Vehicle”, “Move”, “Wait”, etc) with a total of 544,777 event sequences spread across 28,682 buildings. Ground truth labels are available on the 10 different professions of all the individuals. 23 out of the 4,623 individuals are suspected to be part of a terror group.

Given this large database, we show that our augmented BoW can successfully

Table 3: McNemar’s tests on statistical significance between the different methods on the 2 multi-class classification problems. Each column compares two methods. **Left:** Comparing the methods in Figure 7 for the WAAS dataset; **Right:** Comparing the methods in Table 1 for the “knowledge of procedure” skill in the surgery dataset (the other 6 skill classifications were also statistically significant, but are not shown due to space constraints).

	M1 vs M2	M1 vs M3		M1 vs M2	M1 vs M3
χ^2	165.09	530.35	χ^2	4.76	9.33
p -value	<0.0001	0.0026	p -value	0.0291	0.0023

classify people’s professions and detect some of the suspect individuals based on the temporal and structural similarities in their activities. Classification accuracies and confusion matrices are shown in Figure 7. Note that, with our encoding, more than a third of the suspect group are correctly classified which baseline methods failed to capture. This successful identification of suspicious behavior is especially remarkable since those suspects aim for imitating “normal“ behavior and thus their activities are very similar to harmless activities.

2.4.5 Test for Statistical Significance

With McNemar’s chi-square test (with Yates’ continuity correction), we check for the statistical significance between the results of our two multi-class classification problems (Figure 7 and Table 1). For the surgery dataset, though all the 7 skill classifications were statistically significant, due to space constraints, only results on “knowledge of procedure” classification is presented.

The null hypothesis is that the improvements are due to chance. However, as shown in Table 3 for both the datasets, the χ^2 values are greater than the critical value (at 95% significance level) of 3.84 and the p -values are less than the significance level (α) of 0.05. Thus, the null hypothesis can be rejected and we can conclude that the improvements obtained with our methods are statistically significant.

2.4.6 Evaluation Strategy

We evaluated our approach on four diverse classification tasks and used the evaluation strategies that were relevant to the application. Supervised classification was used to evaluate our understanding of vehicle behaviors, detection of skill levels in surgeons and detect anomalies in large surveillance datasets. Results were presented in the form of ROC curves and classification percentages. For functional categorization of Soccer players, we used unsupervised learning to cluster the players into similar groups based on their functionality on the field and presented the results in terms of cluster quality.

While the evaluation strategy in this chapter is based on classic ML approaches such as supervised and unsupervised learning, we will see in the later chapters that we adapt our learning and evaluation strategy to suit the application and the dataset in-hand.

2.5 Conclusion

Spatio-temporal context using BoW models are a promising approach to real-world activity recognition problems where only little is known a-priori about the underlying structure of the data to be analyzed. We presented a significant extension to BoW-based activity recognition, where we augment BoW with temporal information and with both local and global structural information, using temporal encoding, n -grams and randomly sampled regular expressions, respectively.

We showed that, in addition to generally improved activity recognition, our approach also detects skill levels and anomalies in the data, which is important, for example in human behavior analysis applications. We have demonstrated the capabilities of our approach on real-world vision problems and on massive wide-area surveillance simulations.

CHAPTER III

LEVERAGING EGOCENTRIC CONTEXT

In this chapter, we present a technique that leverages the egocentric context by using images, videos and sensor data taken from first-person point-of-view devices to perform egocentric field-of-view (FOV) localization [9]. We define egocentric FOV localization as capturing the visual information from a person’s field-of-view in a given environment and transferring this information onto a reference corpus of images and videos of the same space, hence determining what a person is attending to. Our method matches images and video taken from the first-person perspective with the reference corpus and refines the results using the first-person’s head orientation information obtained using the device sensors. We demonstrate single and multi-user egocentric FOV localization in different indoor and outdoor environments with applications in augmented reality, event understanding and studying social interactions.

3.1 Introduction

A key requirement in the development of interactive computer vision systems is modeling the user, and one very important question is “*What is the user looking at right now?*” From augmented reality to human-robot interaction, from behavior analysis to healthcare, determining the user’s egocentric field-of-view (FOV) accurately and efficiently can enable exciting new applications. Localizing a person in an environment has come a long way through the use of GPS, IMUs and other signals. But such localization is only the first step in understanding the person’s FOV.

The new generation of devices are small, cheap and pervasive. Given that these devices contain cameras and sensors such as gyros, accelerometers and magnetometers, and are Internet-enabled, it is now possible to obtain large amounts of first-person

point-of-view (POV) data unintrusively. Cell phones, small POV cameras such as GoPros, and wearable technology like Google Glass all have a suite of similar useful capabilities. We propose to use data from these first person POV devices to derive an understanding of the user’s egocentric perspective. In this chapter we show results from data obtained with Google Glass, but any other device could be used in its place.

Automatically analyzing the POV data (images, videos and sensor data) to estimate egocentric perspectives and shifts in the FOV remains challenging. Due to the unconstrained nature of the data, no general FOV localization approach is applicable for all outdoor and indoor environments. *Our insight is to make such localization tractable by introducing a reference data-set*, i.e., a visual model of the environment, which is either pre-built or concurrently captured, annotated and stored permanently. All the captured POV data from one or more devices can be matched and correlated against this reference data-set allowing for transfer of information from the user’s reference frame to a global reference frame of the environment. The problem is now reduced from an open-ended data-analysis problem to a more practical data-matching problem. Such reference data-sets already exist; *e.g.*, Google Street View imagery exists for most outdoor locations and recently for many indoor locations. Additionally, there are already cameras installed in many venues providing pre-captured or concurrently captured visual information, with an ever increasing number of spaces being mapped and photographed. Hence there are many sources of visual models of the world which we can use in our approach.

Contributions: We present a method for egocentric FOV localization that directly matches images and videos captured from a POV device with the images and videos from a reference data-set to understand the person’s FOV. We also show how sensor data from the POV device’s IMU can be used to make the matching more efficient and minimize false matches. We demonstrate the effectiveness of our approach across

4 different application domains: (1) egocentric FOV localization in outdoor environments: 250 POV images from different locations in 2 major metropolitan cities matched against the street view panoramas from those locations; (2) egocentric FOV localization in indoor spaces: a 30 minute POV video in an indoor presentation matched against 2 fixed videos cameras in the venue; (3) egocentric video tours at museums: 250 POV images of paintings taken within 2 museums in New York City (Metropolitan Museum of Art and Museum of Modern Art) matched against indoor street view panoramas from these museums (available publicly as part of the Google Art Project [2]); and (4) joint egocentric FOV localization from multiple POV videos: 60 minutes of POV videos captured concurrently from 4 people wearing POV devices at the Computer History Museum in California, matched against each other and against indoor street view panoramas from the museum.

3.2 Related Work

Localization: Accurate indoor localization has been an area of active research [46]. Indoor localization can leverage GSM [91], active badges [127], 802.11b wireless ethernet [61], bluetooth and WAP [3], listeners and beacons [93], radiofrequency [8] technologies and SLAM [66].

Outdoor localization from images or video has also been explored, including methods to match new images to street-side images [98, 134, 99]. Other techniques include urban navigation using a camera mobile phone [95], image geo-tagging based on travel priors [55] and the IM2GPS system [44].

Our approach leverages these methods for visual and sensor data matching with first-person POV systems to determine where the user is attending to.

Egocentric Vision and Attention: Detecting and understanding the salient regions in images and videos has been an active area of research for over three decades. Seminal efforts in the 80s and 90s focused on understanding saliency and attention

from a neuroscience and cognitive psychology perspective [116]. In the late 90s, Illti *et al.* [52] built a visual attention model using a bottom-up model of the human visual system. Other approaches used graph based techniques [43], information theoretical methods [15], frequency domain analysis [51] and the use of higher level cues like face-detection [17] to build attention maps and detect objects and regions-of-interests in images and video.

In the last few years, focus has shifted to applications which incorporate attention and egocentric vision. These include gaze prediction [70], image quality assessment [87], action localization and recognition [101, 27], understanding social interactions [26] and video summarization [65]. Our goal in this work is to leverage image and sensor matching between the reference set and POV sensors to extract and localize the egocentric FOV.

3.3 Egocentric FOV Localization

The proposed methodology for egocentric FOV localization consists of five components: (i) POV data consisting of images, videos and head-orientation information, (ii) a pre-captured or concurrently captured reference dataset, (iii) robust matching pipeline, (iv) match correction using sensor data, and (v) global matching and score computation. An overview of our approach is shown in Figure 8. Each step of the methodology is explained in detail below.

3.3.1 Data collection

POV images and videos along with the IMU sensor data are collected using one or more POV devices to construct a “pov-dataset”. For our experiments, we used a Google Glass. It comes equipped with a 720p camera and sensors such as accelerometer, gyroscope and compass that lets us effectively capture images, videos and sensor data from a POV perspective. Other devices such as cell-phones, which come equipped with cameras and sensors, can also be used.

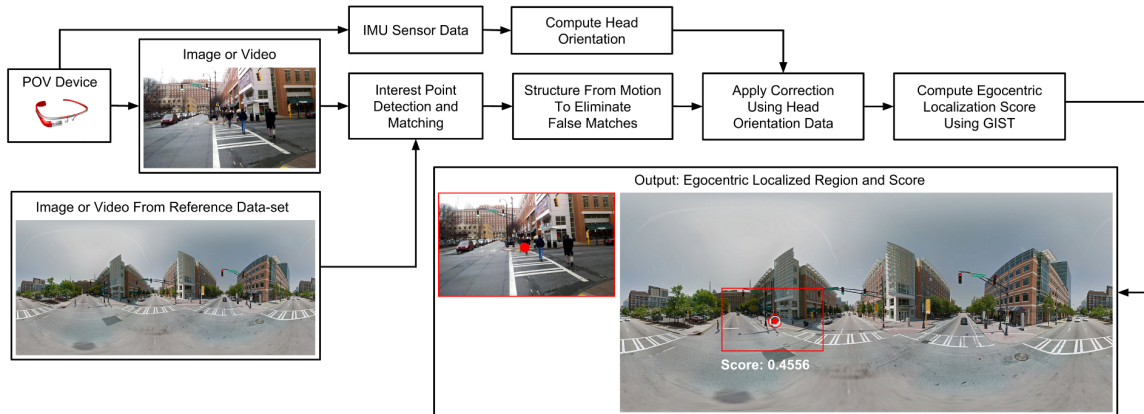


Figure 8: An overview of our egocentric FOV localization system. Given images (or videos) and sensor data from a POV device, and a pre-existing corpus of canonical images of the given location (such as Google street view data), our system localizes the egocentric perspective of the person and determines the person’s region-of-focus.

3.3.2 Reference dataset

A “reference-dataset” provides a visual model of the environment. It can either be pre-captured (and possibly annotated) or concurrently captured (i.e. captured while the person with the POV device is in the environment). Examples of such reference datasets are Google Street View images and pre-recorded videos and live video streams from cameras in indoor and outdoor venues.

3.3.3 Matching

Given the person’s general location, the corresponding reference image is fetched from the reference-dataset using location information (such as GPS) and is matched against all the POV images taken by the person at that location. Since the camera is egocentric, the captured image provides an approximation of the person’s FOV. The POV image and the reference image are typically taken from different viewpoints and under different environmental conditions which include changes in scale, illuminations, camera intrinsics, occlusion and affine and perspective distortions. Given the “in-the-wild” nature of our applications and our data, our matching pipeline is

designed to be robust to these changes.

In the first step of the matching pipeline, reliable interest points are detected both in the POV image, I_{pov} , and the reference image, I_{ref} using maximally stable extremal regions (MSER). The MSER approach was originally proposed by [77], by considering the set of all possible thresholdings of an image, I , to a binary image, I_B , where $I_B(x)=1$ if $I(x) \geq t$ and 0 otherwise. The area of each connected component in I_B is monitored as the threshold is changed. Regions whose rates of change of area with respect to the threshold are minimal are defined as maximally stable and are returned as detected regions. The set of all such connected components is the set of all extremal regions. The word extremal refers to the property that all pixels inside the MSER have either higher (bright extremal regions) or lower (dark extremal regions) intensity than all the pixels on its outer boundary. The resulting extremal regions are invariant to both affine and photometric transformations. A comparison of MSER to other interest point detectors has shown that MSER outperforms the others when there is a large change in viewpoint [81]. This is a highly desirable property since I_{pov} and I_{ref} are typically taken from very different viewpoints. Once the MSERs are detected, standard SIFT descriptors are computed and the correspondences between the interest points are found by matching them using a KD tree, which supports fast indexing and querying.

The interest point detection and matching process may give us false correspondences that are geometrically inconsistent. We use RANSAC (random sample consensus) [29] to refine the matches and in turn eliminate outlier correspondences that do not fit the estimated model. In the final step, the egocentric focus-of-attention is transferred from I_{pov} to I_{ref} . Using three of the reliable match points obtained after RANSAC, the affine transformation matrix, \mathbf{A} , between I_{pov} and I_{ref} is computed. The egocentric focus-of-attention \mathbf{f}_{pov} is chosen as the center of I_{pov} (the red dot in Figure 8). This is a reasonable assumption in the absence of eye-tracking data. The

focus-of-attention, \mathbf{f}_{ref} , in I_{ref} , is given by $\mathbf{f}_{\text{ref}} = \mathbf{A}\mathbf{f}_{\text{pov}}$.

3.3.4 Correction using sensor data

The POV sensor data that we have allows us to add an additional layer of correction to further refine the matches. Modern cellphones and POV devices like Glass come with a host of sensors like accelerometers, gyroscopes and compasses and they internally perform sensor fusion to provide more stable information. Using sensor fusion, these devices report their absolute orientation in the world coordinate frame as a 3×3 rotation matrix R . By decomposing R , Euler angles ψ (yaw), θ (pitch), ϕ (roll) can be obtained. Since Glass is capturing sensor data from a POV perspective, the Euler angles give us the head orientation information, which can be used to further refine the matches. For example, consider a scenario where the user is looking at a high-rise building that has repetitive patterns (such as rectangular windows), all the way from bottom to the top. The vision-based matching gives us a match at the bottom of the building, but the head orientation information suggests that the person is looking up. In such a scenario, a correction can be applied to the match region to make it compatible with the sensor data.

Projecting the head orientation information onto I_{ref} , gives us the egocentric focus-of-attention, \mathbf{f}_{s} , as predicted by the sensor data. The final egocentric FOV localization is computed as: $\mathbf{f} = \alpha\mathbf{f}_{\text{s}} + (1 - \alpha)\mathbf{f}_{\text{ref}}$, where α is a value between 0 and 1 and is based on the confidence placed on the sensor data. Sensor reliability information is available in most of the modern sensor devices. If the device sensors are unreliable then α is set to a small value. Relying solely on either vision based matching or on sensor data is not a good idea. Vision techniques fail when the images are drastically different or have fewer features and sensors tend to be noisy and the readings drift over time. We found that first doing the vision based matching and then applying a α -weighted correction based on the sensor data gives us the best of

both worlds.

3.3.5 Global Matching and Score Computation

We now have a match window that is based on reliable MSER interest point detection followed by SIFT matching and RANSAC based outlier rejection and sensor based correction. Although this match window is reliable, it is still based only on local features without any global context of the scene. There are several scenarios in the real world (like urban environments), where we have repetitive and commonly occurring patterns and local features that may result in an inaccurate match window. In this final step, we do a global comparison and compute the egocentric localization score.

Global comparison is done by comparing the match window, W_{ref} located around \mathbf{f}_s in I_{ref} , with I_{pov} (*i.e.*, the red match windows of the bottom image in Figure 8). This comparison is done using global GIST descriptors [90]. A GIST descriptor gives a global description of the image based on the image’s spectral signatures and tells us how visually similar the two images are. GIST descriptors \mathbf{q}_{pov} and \mathbf{q}_{ref} are computed for I_{pov} and W_{ref} respectively and final egocentric FOV localization score is computed as the $L2$ -distance between the GIST descriptors: $\| \mathbf{q}_{pov} - \mathbf{q}_{ref} \| = \sqrt{(\mathbf{q}_{pov} - \mathbf{q}_{ref}) \cdot (\mathbf{q}_{pov} - \mathbf{q}_{ref})}$. Scoring quantifies the confidence in our matches and by thresholding on the score, we can filter out incorrect matches.

3.4 Applications and Results

To evaluate our approach and showcase different applications, we built 4 diverse datasets that include both images and videos in both indoor and outdoor environments. All the POV data was captured with a Google Glass.

3.4.1 Outdoor Urban Environments

Egocentric FOV localization in outdoor environments has applications in areas such as tourism, assistive technology and advertising. To evaluate our system, 250 POV images (of dimension 2528x1856) along with sensor data (roll, pitch and yaw of the head) was captured at different outdoor locations in two major metropolitan cities. The reference dataset consists of the 250 street view panoramas (of dimension 3584x1536) from those locations. Based on the user’s GPS location, the appropriate street view panorama was fetched and used for matching. Ground truth was provided by the user who documented his point of attention in each of the 250 POV images. However we have to take into account the fact that we are only tracking the head orientation using sensors and not tracking the eye movement. Humans may or may not rotate their heads completely to look at something; instead they may rotate their head partially and just move their eyes. We found that this behavior (of keeping the head fixed while moving the eyes) causes a circle of uncertainty of radius R around the true point-of-attention in the reference image. To calculate its average value, we conducted a user-study with 5 participants. The participants were instructed to keep their heads still and use only their eyes to see as far to the left and to the right as they could without the urge to turn their heads. This mean radius of their natural eye movement was measured to be 330 pixels for outdoor urban environments. Hence for our evaluation we consider the egocentric FOV localization to be successful if the estimated point-of-attention falls within a circle of radius $R = 330$ pixels around the ground truth point-of-attention.

Experimental results show that without using sensor data, egocentric FOV localization was accurate in 191/250 images for a total accuracy of 76.4%. But when sensor data was included, the accuracy rose to 92.4%. Figure 9 shows the egocentric FOV localization results and the shifts in FOV over time. Discriminative objects such as landmarks, street signs, graffiti, logos and shop names helped in the getting good

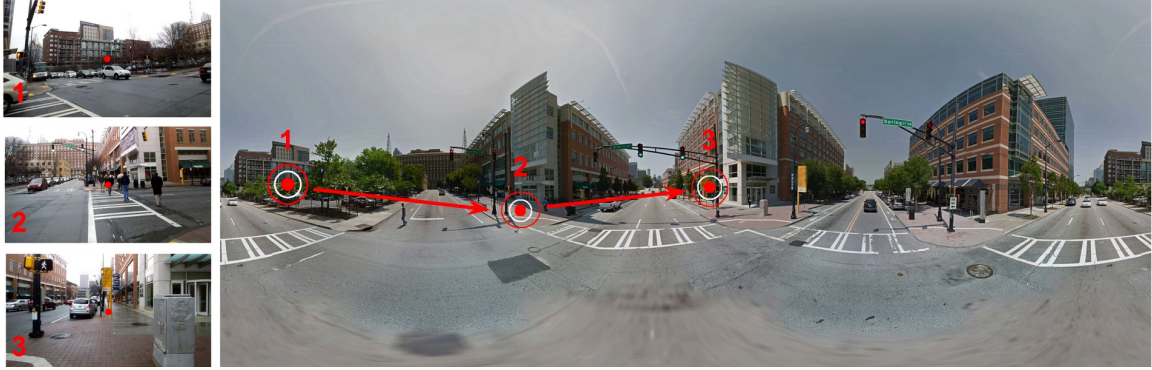


Figure 9: Egocentric FOV localization in outdoor environments. The images on the left are the POV images taken from Glass. The red dot shows the focus-of-attention. The panorama on the right shows the localization (target symbols) and the shifts in the FOV over time (red arrows). Note the change in season and pedestrian traffic between the POV images and the reference image.

matches. Repetitive and commonly occurring patterns like windows and vegetation caused initial failures but most of them were fixed when the sensor correction was applied.

3.4.2 Presentations in Indoor Spaces

There are scenarios where a pre-built reference dataset (like street view) is not available for a given location. This is especially true for indoor environments that have not been as thoroughly mapped as outdoor environments. In such scenarios, egocentric FOV localization is possible with a reference dataset that is concurrently captured along with the POV data. To demonstrate this, a 30 minute POV video along with sensor data was captured during an indoor presentation. The person wearing Glass was seated in the audience in the first row. The POV video is 720p at 30 fps. The reference dataset consists of videos from two fixed cameras at the presentation venue. One camera was capturing the presenter while the other camera was pointed at the audience. The reference videos are 1080p at 30 fps. Ground truth annotations for every second of the video were provided by the user who wore Glass and captured the



Figure 10: Egocentric FOV localization in indoor environments. The images on the first column show the room layout. The presenter is shown in Green and the person wearing Glass is shown in Blue with his egocentric view shown by the blue arrow. The second column shows the POV video frames from Glass. The red dot shows the focus-of-attention. The third and fourth column show the presenter cam and the audience cam respectively. The localization is shown by the target symbol and the selected camera is shown by the red bounding box. The person wearing Glass is highlighted by the blue circle in the presenter camera views.

POV video. So, we have $60 \times 30 = 1800$ ground truth annotations. As with the previous dataset, we empirically estimated R to be 240 pixels. Experimental results show that egocentric FOV localization and camera selection was accurate in 1722/1800 cases for a total accuracy of 95.67%. Figure 10 shows the FOV localization and camera selection results.

3.4.3 Egocentric Video Tours in Museums

Public spaces like museums are ideal environments for an egocentric FOV localization system. Museums have exhibits that people explicitly pay attention to and want to learn more about. Similar to audio-tours that are available in museums, we demonstrate a system for attention-driven egocentric video tours. Unlike in an audio tour where a person has to enter the exhibit number to hear details about it, our video tour system recognizes the exhibit when the person looks at it and brings up a cue card on the wearable device giving more information about the exhibit.

For our evaluation, we captured 250 POV images of paintings at 2 museums in

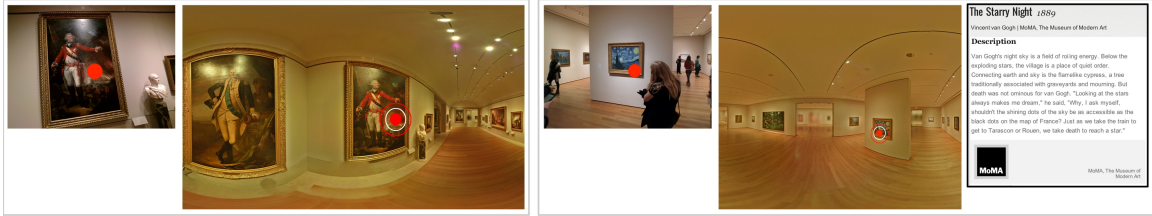


Figure 11: Egocentric FOV localization in indoor art installations. The images on the left are the POV images taken from Glass. The red dot shows the focus-of-attention. The images to their right are panoramas from indoor streetview that correctly shows the localization result (target symbol). When available, the details of the painting are shown. This information is automatically fetched, using the egocentric FOV location as the cue. For the painting on the right (Van Gogh’s “The Starry Night”), an information card shows up and provides information about the painting.

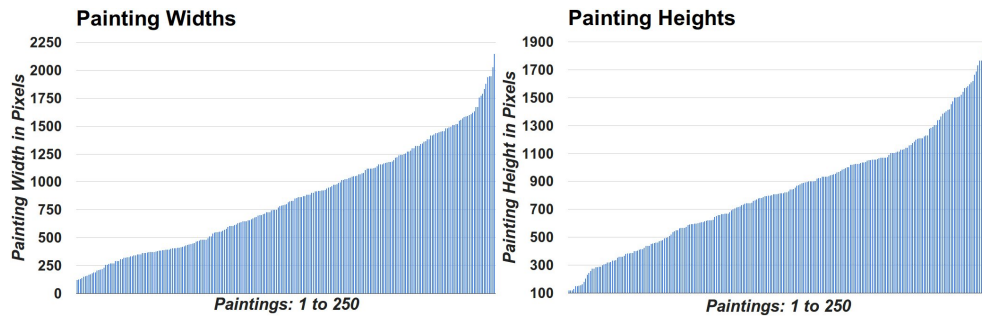


Figure 12: The widths and heights of the 250 paintings, sorted in ascending order based on their value. We can see that our dataset has a good representation of paintings of varying widths and heights.

New York City - The Metropolitan Museum of Art and The Museum of Modern Art. The reference dataset consists of indoor street view panoramas from these museums, made available as part of the Google Art Project [2]. Since this dataset consists of paintings, which have a fixed structure (a frame enclosing the artwork), we have a clear definition of correctness: egocentric FOV localization is deemed to be correct if the estimated focus-of-attention is within the frame of the painting in the reference image. Experimental results show that the localization was accurate in 227/250 images for a total accuracy of 90.8%. Figure 12 shows the distribution of the widths and heights of the paintings in our dataset. We can see that paintings of all widths and heights are well represented.

The Google Art panoramas are annotated with information about the individual paintings. On successful FOV localization, we fetch the information on the painting that the person is viewing and display it on Glass or as an overlay. Figure 11 shows the FOV localization results and the painting information that was automatically fetched and shown on Glass.

3.4.4 Joint Egocentric FOV Localization

When we have a group of people wearing POV devices within the same event space, egocentric FOV localization becomes much more interesting. We can study joint FOV localization (i.e. when two or more people are simultaneously attending to the same object), understand the social dynamics within the group and gather information about the event space itself.

Joint FOV localization can be performed by matching the videos taken from one POV device with the videos taken from another POV device. If there are n people in the group, $P = \{p_i | i \in [1, n]\}$, then we have n POV videos: $V = \{v_i | i \in [1, n]\}$. In the first step, all the videos in V are synchronized by time-stamp. In the second step, k videos (where $k \leq n$) are chosen from V and matched against each other, which results in a total of $\binom{n}{k}$ matches. Matching is done frame-by-frame, by treating frame from one video as I_{pov} and the frames from the other videos as I_{ref} . By thresholding the egocentric FOV localization scores, we can discover regions in time when the k people were jointly paying attention to the same object. Finally, in the third step, the videos can be matched against the reference imagery from the event space to find out *what* they were jointly paying attention to.

We conducted our experiments with $n = 4$ participants. The 4 participants wore Glass and visited the Computer History Museum in California. They were instructed to behave naturally, as they would on a group outing. They walked around in the museum looking at the exhibits and talking with each other. A total of 60 minutes

of POV videos and the corresponding head-orientation information were captured from their 4 Glass devices. The videos are 720p at 30fps. The reference dataset consists of indoor street view panoramas from the museum. Next, joint egocentric FOV localization was performed by matching pairs of videos against each other, i.e. $k = 2$, for a total of 6 pairs of matches. Figure 14 shows the results for 25,000 frames of video for all the 6 match pairs. The plot shows the instances in time when groups of people were paying attention to the same exhibit. Furthermore, we get an insight into the social dynamics of the group. For example, we can see that P2 and P3 were moving together but towards the end P3 left P2 and started moving around with P1. Also, there are instances in time when all the pairs of videos match which indicates that the group came together as a whole. One such instance is highlighted in Figure 14 by the orange vertical line. There are also instances when the 4 people split into two groups. This is shown by the green vertical line in Figure 14.

Joint egocentric FOV localization also helps us get a deeper understanding of the event space. Interesting exhibits tend to bring people together for a discussion and result in higher joint egocentric attention. It is possible to infer this from the data by matching the videos with the reference images and labeling each exhibit with the number of people who jointly viewed it. By overlaying the exhibits on the floorplan, we can generate a heat map of the exhibits where hotter regions indicate more interesting exhibits that received higher joint attention. This is shown in Figure 13. Getting such an insight has practical applications in indoor space planning and the arrangement and display of exhibits in museums and other similar spaces.

3.4.5 Evaluation Strategy

In the previous chapter we used classic ML evaluation methodologies such as supervised and unsupervised learning on the activity recognition and anomaly detection datasets. However, in this chapter, the evaluation strategy is different due to the

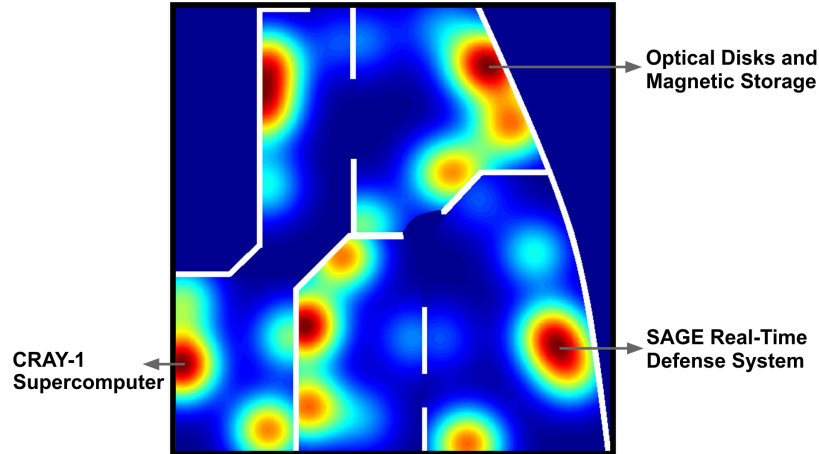


Figure 13: A heatmap overlaid on a section of the Computer History Museum’s floorplan. Hotter regions in the map represent exhibits which had joint egocentric attention from more people. Three of the hottest regions are labeled to show the underlying exhibits that brought people together and probably led to further discussions among them.

nature of our applications and the datasets that we use. Here the ground-truth is the users’ POV when the image or video was captured using the POV device. Given this ground-truth information, we account for the mean radius of natural eye movements and deem the localization to be accurate if the estimated point-of-attention falls within this radius around the ground-truth point-of-attention. Furthermore, some of our results are subjective in nature and can only be visualized pictorially. For example, the attention heatmap of the museum generated using the joint-attention information was shown as a visualization in Figure 13. Since this result is first of its kind we do not have ground-truth data for this, and in-turn, this makes the result entirely subjective by nature.

3.5 Discussion

One of the assumptions here is the availability of reference images in indoor and outdoor spaces. This may not be true for all situations. Also, it may not be possible to capture reference data concurrently (as in the indoor presentation dataset) due to restrictions by the event managers and/or privacy concerns. However, our assumption

does hold true for a large number of indoor and outdoor spaces which makes the proposed approach practical and useful.

There are situations where the proposed approach may fail. While our matching pipeline is robust to a wide variation of changes in the images, it may still fail if the reference image is drastically different from the POV image (for example, a POV picture taken in summer matched against a reference image taken on a white snowy winter). Another reason for failure could be when the reference dataset is outdated. In such scenarios, the POV imagery will not match well with the reference imagery. However these drawbacks are only temporary. With the proliferation of cameras and the push to map and record indoor and outdoor spaces, reference data for our approach will only become more stable and reliable.

Our reference images are 2D models of the scene (for example, Street View panoramas). Moving to 3D reference models could provide a more comprehensive view of the event space and result in better FOV localization. But this would require a computationally intensive matching pipeline which involves 2D to 3D alignment and pose estimation.

3.6 Conclusion

In this chapter, we demonstrated a working system that can effectively leverage egocentric context and localize egocentric FOVs, determine the person’s point-of-interest, map the shifts in FOV and determine joint attention in both indoor and outdoor environments from one or more POV devices. Several practical applications were presented on “in-the-wild” real-world datasets.



Figure 14: The plot on the top shows the joint egocentric attention between groups of people. The x-axis shows the progression of time, from frame 1 to frame 25,000. Each row shows the result of joint egocentric FOV localization, i.e. the instances in time when pairs of people were jointly paying attention to the same exhibit in the museum. The orange vertical line indicates an instance in time when all the people (P1, P2, P3 and P4) were paying attention to the same exhibit. The green vertical line indicates an instance in time when P1 and P4 were jointly paying attention to an exhibit while P2 and P3 were jointly paying attention to a different exhibit. The corresponding frames from their Glass videos is shown. When matched to the reference street view images, we can discover the exhibits that the groups of people were viewing together and were probably having a discussion about. Details of the exhibit was automatically fetched from the reference dataset's annotation.

CHAPTER IV

LEVERAGING GEOGRAPHIC CONTEXT

In this chapter, we look at how geographic context can be leveraged to make challenging “in-the-wild” object recognition tasks more tractable using the problem of food recognition in restaurants as a case-study [11]. The pervasiveness of mobile cameras has resulted in a dramatic increase in food photos, which are pictures reflecting what people eat. In this chapter, we study how taking pictures of what we eat in restaurants can be used for the purpose of automating food journaling. We propose to leverage the context of where the picture was taken, with additional information about the restaurant, available online, coupled with state-of-the-art computer vision techniques to recognize the food being consumed. To this end, we demonstrate image-based recognition of foods eaten in restaurants by training a classifier with images from restaurant’s online menu databases. We evaluate the performance of our system in unconstrained, real-world settings with food images taken in 10 restaurants across 5 different types of food (American, Indian, Italian, Mexican and Thai).

4.1 Introduction

Recent studies show strong evidence that adherence to dietary self-monitoring helps people lose weight and meet dietary goals [16]. This is critically important since obesity is now a major public health concern associated with rising rates of chronic disease and early death [57].

Although numerous methods have been suggested for addressing the problem of poor adherence to nutrition journaling [4, 97, 133], a truly practical system for objective dietary monitoring has not yet been realized; the most common technique for logging eating habits today remains self-reports through paper diaries and more

recently, smartphone applications. This process is tedious, time-consuming, prone to errors and leads to selective under reporting [34].

While needs for automated food journaling persist, we are seeing an ever increasing growth in people photographing what they eat. In this chapter we present a system and approach for automatically recognizing foods eaten at restaurants from first-person food photos with the goal of facilitating food journaling. The methodology we employ is unique because it leverages sensor data (i.e., location) captured at the time photos are taken. Additionally, online resources such as restaurant menus and online images are used to help recognize foods once a location has been identified.

Our motivation for focusing on restaurant eating activities stems from findings from recent surveys indicating a trend towards eating out versus eating at home. In 1970, 25.9 percent of all food spending was on food away from home; by 2012, that share rose to its highest level of 43.1 percent [119]. Additionally, 8 in 10 Americans report eating at fast-food restaurants at least monthly, with almost half saying they eat fast food at least weekly [33].

Research in the computer vision community has explored the recognition of either a small sub-set of food types in controlled laboratory environments [19, 132] or food images obtained from the web [47]. However, there have been only a few validated implementations that address the challenge of food recognition from images taken “in the wild” [56]. Systems that rely on crowdsourcing, such PlateMate [88], have shown promise but are limited in terms of cost and scalability. Additionally, privacy concerns might arise when food photographs are reviewed by untrusted human computation workers [110].

In this chapter, we seek an approach that supports automatic recognition of food, leveraging the context of where the photograph was taken. Our contributions are:

- An automatic workflow where online resources are queried with contextual sensor data to find food images and additional information about the restaurant

where the food picture was taken, with the intent to build classifiers for food recognition.

- An image classification approach using the SMO-MKL multi-class SVM classification framework with features extracted from test photographs.
- An in-the-wild evaluation of our approach with food images taken in 10 restaurants across 5 different types of cuisines (American, Indian, Italian, Mexican and Thai).
- A comparative evaluation focused on the effect of location data in food recognition results.

In this chapter, we concentrate on food recognition, leveraging the additional context that is available (location, websites, etc.). Our goal here is to in essence, using food and restaurants as the domain, demonstrate the value of external context, coupled with image recognition to support classification. We believe that the same method can be used for many other domains.

4.2 Related Work

Various sensor-based methods for automated dietary monitoring have been proposed over the years. Amft and Troster [4] explored sensors in the wrists, head and neck to automatically detect food intake gestures, chewing, and swallowing from accelerometer and acoustic sensor data. Sazonov et al. built a system for monitoring swallowing and chewing using a piezoelectric strain gauge positioned below the ear and a small microphone located over the laryngopharynx [97]. Yatani and Truong presented a wearable acoustic sensor attached to the user’s neck [133] while Cheng et al. explored the use of a neckband for nutrition monitoring [22].

With the emergence of low-cost, high-resolution wearable cameras, recording individuals as they perform everyday activities such as eating has been gaining appeal

[5]. In this approach, individuals wear cameras that take first-person point-of-view photographs periodically throughout the day. Although first-person point-of-view images offer a viable alternative to direct observation, one of the fundamental problems is image analysis. All captured images must be manually coded for salient content (e.g., evidence of eating activity), a process tends to be tedious and time-consuming.

Over the past decade, research in computer vision is moving towards “in the wild” approaches. Recent research has focussed on recognizing realistic actions in videos [73], unconstrained face verification and labeling [58] and objection detection and recognition in natural images [25]. Food recognition in the wild using vision-based methods is growing as a topic of interest, with Kitamura et al. [56] showing promise.

Finally, human computation lies in-between completely manual and fully-automated vision-based image analysis. PlateMate [88] crowdsources nutritional analysis from food photographs using Amazon Mechanical Turk, and Thomaz et al. investigated the use of crowdsourcing to detect [111] eating moments from first-person point-of-view images. Despite the promise of these crowdsourcing-based approaches, there are clear benefits to a fully automated method in economic terms, and possibly with regards to privacy as well.

4.3 Methodology

Recognizing foods from photographs is a challenging undertaking. The complexity arises from the large number of food categories, variations in their appearance and shape, the different ways in which they are served and the environmental conditions they are presented in. To offset the difficulty of this task, the methodology we propose (Figure 15) centers on the use of location information about the eating activity, and also restaurant menu databases that can be queried online. As noted, our technique is specifically aimed at eating activities in restaurants as we leverage the context of restaurant related information for classification.

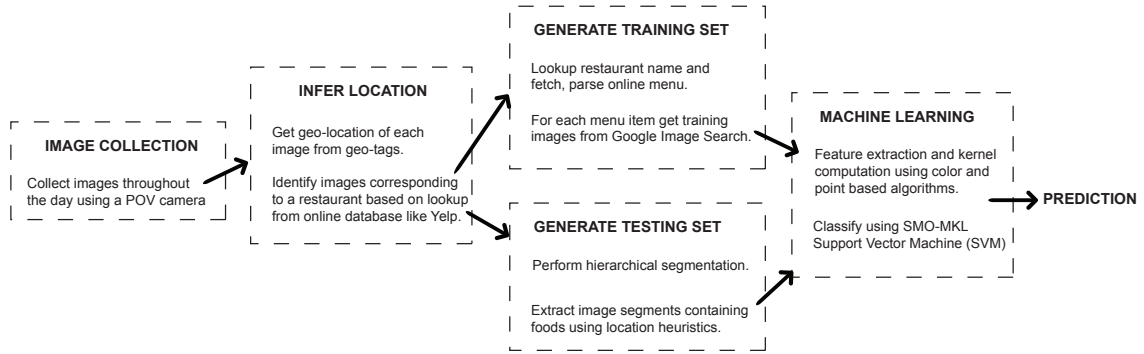


Figure 15: An overview of our automatic food recognition approach.

4.3.1 Image Acquisition

The first step in our approach involves the acquisition of food images. The popularity of cameras in smartphones and wearable devices like Google Glass makes it easy to capture food images in restaurants. In fact, many food photographs communities such as FoodGawker have emerged over the last several years, all centered on food photo sharing. Photographing food is also hitting major photo sharing sites like Instagram, Pinterest and Flickr, and food review sites like Yelp. These food-oriented photo activities illustrate the practicality of using manually-shot food photos for food recognition.

4.3.2 Geo-Localizing Images

The second step involves associating food photos with longitude and latitude coordinates. If the camera that is being used supports image geo-tagging, then the process of localizing images is greatly simplified. Commodity smart-phones and cameras like the Contour and SenseCam come with built-in GPS capabilities. If the geo-tag is not available, image localization techniques can be used [134]. Once location is obtained for all captured images, the APIs of Yelp and Google Places are valuable for matching the images' geo-tags coincide with the geo-tag of a restaurant.



Figure 16: Weakly-labeled training images obtained from Google Image search for 3 classes of food: **Left:** Basil Fried Rice; **Center:** Curry Katsu; **Right:** Lo Mein.

4.3.3 Weakly Supervised Learning

Being able to localize images to a restaurant greatly constrains the problem of food classification in the wild. A strong assumption can be made that the food present in the images must be from one of the items on the restaurant’s menu. This key observation makes it possible to build a weakly supervised classification framework for food classification. The subsequent sections describe in detail the gathering of weakly-labeled training data, preparing the test data and classification using the SMO-MKL multi-class SVM classification framework [121].

4.3.3.1 Gathering Training Data

We start with collecting images localized to a particular restaurant R . Once we know R , we can use the web as a knowledge-base and search for R ’s menu. This task is greatly simplified thanks to online data-sources like Yelp, Google Places, Allmenus.com and Openmenu.com, which provides comprehensive databases of restaurant menus.

Let the menu for R be denoted by M_R and let the items on the menu be m_i . For each $m_i \in M_R$, the top 50 images of m_i are downloaded using search engines like Google Image search. This comprises the weakly-labeled training data. Three examples are shown in Figure 16. From the images, it is possible to see that there is a high degree of intra-class variability in terms of color and presentation of food. As

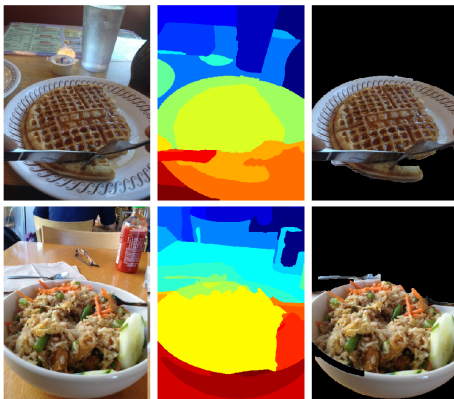


Figure 17: Extracting segments using hierarchical segmentation. The final segmented image is shown on the right.

is the case with any state-of-the-art object recognition system, our approach relies on the fact that given sufficient number of images for each class, it should be possible to learn common patterns and statistical similarities from the images.

4.3.3.2 Preparing Testing Data

The test images, localized to restaurant R , are segmented using hierarchical segmentation and the segments are extracted from parts of the image where we expect the food to be present [6]. The final set of segmented images forms our test data. An example is shown in Figure 17.

4.3.3.3 Feature Descriptors

Choosing the right combination of feature detectors, descriptors and classification backend is key to achieving good accuracy in any object recognition or image categorization task. While salient point detectors and corresponding region descriptors can robustly detect regions which are invariant to translation, rotation and scale [74, 80], illumination changes can still cause performance to drop. This is a cause of concern when dealing with food images, since images taken at restaurants are typically indoors and under varying lighting conditions. Recent work by van de Sande et al. [120] systematically studies the invariance properties and distinctiveness of color descriptors.

The results of this study guided the choice of the descriptors in our approach. For the classification back-end, we use Multiple Kernel Learning (MKL), which in recent years, has given robust performance on object categorization tasks [7, 105, 121].

For feature extraction from the training and test data, a Harris-Laplace point detector is used since it has shown good performance for category recognition tasks [135] and is scale-invariant. However the choice of feature descriptor is more complicated. As seen in Figure 16, there is a high degree of intra-class variability in terms of color and lighting. Based on the recent work by van de Sande et al. [120] that studies the invariance properties and distinctiveness of various color descriptors on light intensity and color changes, we pick the following six descriptors, 2 color-based and 4 SIFT-based (Scale-Invariant Feature Transform [74]):

- Color Moment Invariants: Generalized color moments M_{pq}^{abc} (of order $p + q$ and degree $a + b + c$) have been defined as $M_{pq}^{abc} = \int \int x^p y^q [I_R(x, y)]^a [I_G(x, y)]^b [I_B(x, y)]^c dx dy$. Color moment invariants are those combinations of generalized color moments that allow for normalization against photometric changes and are invariant to changes and shifts in light intensity and color.
- Hue Histograms: Based on the observation that the certainty of hue is inversely proportional to the saturation, each hue sample in the hue histogram is weighted by its saturation. This helps overcome the (known) instability of hue near the gray axis in HSV space. The descriptors obtained are invariant to changes and shifts in light intensity.
- C-SIFT: The descriptors are built using the C-invariant (normalized opponent color space). C-SIFT is invariant to changes in light intensity.
- OpponentSIFT: All the channels in the opponent color space are described using SIFT descriptors. They are invariant to changes and shifts in light intensity.

- RGB-SIFT: SIFT descriptors are computed for every RGB channel independently. The resulting descriptors are invariant to changes and shifts in light intensity and color.
- SIFT: The original SIFT descriptor proposed by Lowe [74]. It is invariant to changes and shifts in light intensity.

4.3.3.4 Classification Using SMO-MKL

For a given restaurant R , 100,000 interest points are detected in the training data and for each of the 6 descriptors, visual codebooks are built using k -means clustering with $k = 1000$. Using these codebooks, bag-of-words (BoW) histograms are built for the training images. Similarly, interest points are detected in the test images and BoW are built for the 6 descriptors (using the visual codebooks generated with the training data).

For each of the 6 sets of BoW features, extended Gaussians kernels of the following form are computed:

$$K(H_i, H_j) = \exp\left(-\frac{1}{A}D(H_i, H_j)\right) \quad (2)$$

where $H_i = \{h_{in}\}$ and $H_j = \{h_{jn}\}$ are the BoW histograms (scaled between 0 to 1 such that they lie within a unit hypersphere) and $D(H_i, H_j)$ is the χ^2 distance defined as

$$D(H_i, H_j) = \frac{1}{2} \sum_{n=1}^V \frac{(h_{in} - h_{jn})^2}{h_{in} + h_{jn}} \quad (3)$$

where V is the vocabulary size (1000, in our case). The parameter A is the mean value of the distances between all the training examples [135]. Given the set of these N base kernels $\{K_k\}$ (in our case $N = 6$), linear MKL aims to learn a linear combination of the base kernels: $K = \sum_{k=1}^N \alpha_k K_k$

But the standard MKL formulation subject to l_1 regularization leads to a dual that is not differentiable. Hence the Sequential Minimal Optimization (SMO) algorithm cannot be applied and more expensive alternatives have to be pursued. Recently, Vishwanathan et al. showed that it is possible to use the SMO algorithm if the focus is on training p -norm MKL, with $p > 1$ [121]. They also show that the SMO-MKL algorithm is robust and significantly faster than the state-of-the-art p -norm MKL solvers. In our experiments, we train and test using the SMO-MKL SVM.

4.4 Study & Evaluation

We perform two sets of experiments to evaluate our approach. In the first set of experiments, we compare the feature extraction and classification techniques used in this chapter, with the state-of-the-art food recognition algorithms on the PFID benchmark data-set [19]. This validates our proposed approach. In the second set of experiments, we measure the performance of the proposed approach for “in-the-wild” food recognition.

4.4.1 Comparative Evaluations

We study the performance of the 6 feature descriptors and SMO-MKL classification on the PFID food data-set. The PFID dataset is a collection of 61 categories of fast food images acquired under lab conditions. Each category contains 3 different instances of food with 6 images from 6 view-points in each instance. In order to compare our results with the previous published results on PFID [19, 132], we follow the same protocol used by them, i.e. a 3-fold cross-validation is performed with 12 images from one instance being used for training while the other 6 images from the remaining instance are used for testing.

The results of our experiments are shown in Figure 18. MKL gives the best performance and improves the state-of-the-art [132] by more than 20%. It is interesting to note that the SIFT descriptor used in our approach achieves 34.9% accuracy whereas

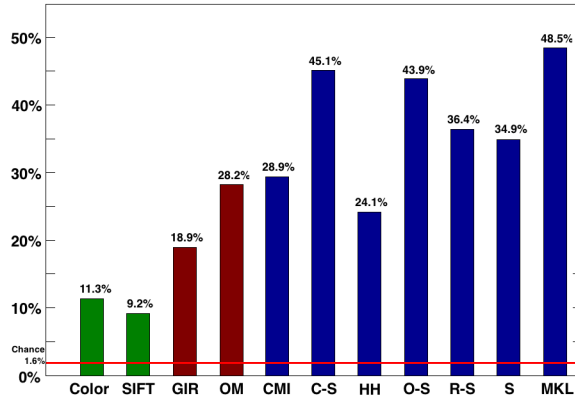


Figure 18: Performance of the 6 feature descriptors and SMO-MKL on the PFID data-set. The first two results (shown in green) are the baseline for PFID published by [19]. The next two (shown in red) are the results obtained by using Global Ingredient Representation (GIR) and Orientation and Midpoint Category (OM) [132]. The rest of the results (in blue) are one ones obtained using the 6 feature descriptors and MKL (CMI: Color Moment Invariant, C-S: C-SIFT, HH: Hue-Histogram, O-S: OpponentSIFT, R-S: RGB- SIFT, S: SIFT and MKL: Multiple Kernel Learning). MKL gives the best performance on this data-set.

the SIFT descriptor used in the PFID baseline [19] achieves 9.2% accuracy. The reason for this difference is that the authors of the PFID baseline use LIB-SVM for classification with its default parameters. However, by switching to the χ^2 kernel (and ensuring that the data is scaled) and by tuning the SVM parameters (through a grid-search over the space of C and γ), we can get a significant boost in performance with just SIFT features alone.

4.4.2 Food Recognition in Restaurants

To study the performance and the practicality of our approach, experiments were conducted on images collected from restaurants across 5 different cuisines: American, Indian, Italian, Mexican and Thai. To discount for user and location bias, 3 different individuals collected images on different days from 10 different restaurants (2 per cuisines). The data collection was done in two phases. In the first phase, the food images were captured using smartphone cameras. In total, 300 “in-the-wild” food images ($5 \text{ cuisines} \times 6 \text{ dishes/cuisine} \times 10 \text{ images/dish}$) were obtained. In the second



Figure 19: Sample (12 out of 600) of the “in-the-wild” images used in testing.

Table 4: Classification results showing the performance of the various feature descriptors on the 5 cuisines. The columns are: CMI: Color Moment Invariant, C-S: C-SIFT, HH: Hue-Histogram, O-S: OpponentSIFT, R-S: RGB-SIFT, S: SIFT and MKL: Multiple Kernel Learning.

	CMI	C-S	HH	O-S	R-S	S	MKL
American	45.8%	51.7%	43.3%	43.3%	37.5%	29.2%	67.5%
Indian	44.2%	74.2%	55.0%	59.2%	69.2%	65.0%	80.8%
Italian	33.3%	52.5%	67.5%	74.2%	66.7%	49.2%	67.5%
Mexican	36.7%	35.8%	20.8%	37.5%	24.2%	33.3%	43.3%
Thai	27.5%	36.7%	25.0%	33.3%	50.8%	30.8%	50.8%

phase, data collection was repeated using a Google Glass and an additional 300 images were captured. These 600 “in-the-wild” images, form our test data-set. A sample of these test images is shown in Figure 19.

Using the geo-location information, the menu for each restaurant was automatically retrieved. For our experiments, we restricted the training to 15 dishes from each cuisine (selected based on online popularity). For each of the 15 dishes on the menu, 50 training images were downloaded using Google Image search. Thus, a total of 3,750 weakly-labeled training images were downloaded (5 cuisines \times 15 menu-items/cuisine \times 50 training-images/menu-item).

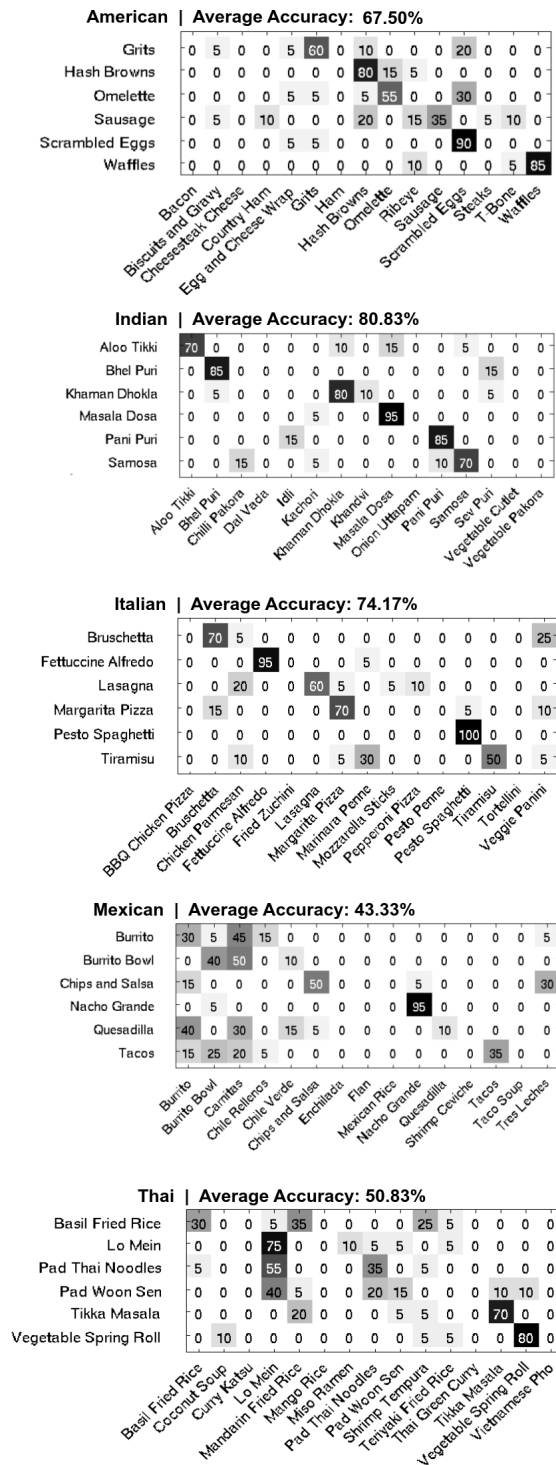


Figure 20: Confusion matrices for the best performing features of Table 4 (for each of the 5 cuisines). Darker colors show better recognition. The 6 food classes in the rows are the ones used for testing and the 15 food classes in the columns are the ones used for training. The overall average accuracy is 63.33%

Next, we perform interest point detection, feature extraction, codebook building for BoW representation, kernel pre-computation and finally classification using SMO-MKL. The results are summarized in Table 4 and the individual confusion matrices are shown in Figure 20. We achieve good classification accuracy with American, Indian and Italian cuisines. However, for the Mexican and Thai cuisines, the accuracy is limited. It could be due to the fact that there is a low degree of visible variability between food types belonging to the same cuisines. For example, in the confusion matrix for Thai, we can see that Basil Fried Rice is confused with Mandarin Fried Rice and Pad Thai Noodles is confused with Lo Mein. It could be very hard, even for humans, to distinguish between such classes by looking at their images.

From Table 4, we can see that there is no single descriptor that works well across all the 5 cuisines. This could be due to the high-degree of variation in the training data. However, combining the descriptors using MKL yields the best performance in 4 out of the 5 cases.

4.4.3 Recognition Without Location Prior

Our approach is based on the hypothesis that knowing the location (through geo-tags) helps us in narrowing down the number of food categories which in turn boosts recognition rates. In order to test this hypothesis, we disregard the location information and train our SMO-MKL classifier on all of the training data (3,750 images). With this setup, accuracy across our 600 test images is 15.67%. On the other hand, the overall average accuracy across the 5 cuisines (from Figure 20) is 63.33%. We can see that the average performance increased by 47.66% when location prior was included. This provides validation that knowing the location of eating activities helps in food recognition, and that it is better to build several smaller restaurant/cuisine specific classifiers rather than one all-category food classifier.

4.4.4 Evaluation Strategy

The evaluation strategies in this thesis are driven by the nature of our application. In the first chapter, we employed a traditional ML approach of supervised and unsupervised learning for activity recognition, skill classification, anomaly detection and functional categorization. In the second chapter, our evaluations included visualizations and a non-traditional approach of figuring out a person’s mean eye-movement radius around the ground-truth point-of attention and using that information to determine the accuracy of our estimated point-of-attention. In this chapter, we evaluate our results within a multi-class semi-supervised classification framework. Unlike the applications in the first chapter, where the training data and the testing data came from the same dataset (test/train split), in this chapter the training data is obtained in a semi-supervised manner from Google Images and the testing data is obtained from the pictures of the food that we are eating. Given that our training data has weak labels, we employ a SMO-MKL SVM learning based approach and report the percentage accuracy of each of our food categories.

4.5 Discussion

In this section we discuss several important points pertaining to the generalizability of our approach, implementation issues, and practical considerations.

4.5.0.1 Generalizability

The automatic food identification approach that we propose is focused on eating activities in restaurants. Although this might seem limiting, eating out has been growing in popularity and 43.1% of food spending was reported to having been spent in foods away from home in 2012 [33, 119]. Moreover, we feel that eating and food information gathered in restaurants is more valuable for dietary self-monitoring than food information obtained at home, since individuals are more likely to know food

types and the composition of food items prepared in their own homes.

We designed our study and evaluation with the goal of maximizing the external validity of our results. We evaluated our approach by having three individuals collect images from the most popular restaurant types by cuisine in the US on different days and using two different devices (smartphones and Google Glass). We feel confident that our methodology will scale in the future, especially since it leverages sensor data, online resources and practices around food imagery that will become increasingly more prevalent in years to come.

One important aspect of the approach is that it depends on weakly-labeled training images obtained from the web. The high-degree of intra-class variability for the same food across different restaurants has a negative effect on performance. A promising alternative is to train on (automatically acquired) food images taken at the same restaurant as the one where the test images were taken. While getting this kind of data seems difficult, it may soon be possible. A recently launched service by Yelp (among others), allows users to upload photos of their food. With such crowd-sourced imagery available for a given restaurant, it may soon be possible to train specialized classifiers for that restaurant. In our future work, we plan to test this hypothesis and improve the recognition accuracies.

4.5.0.2 Location Error

We not only identify the cuisine that the individual is eating, but also identify the specific dish that is being consumed. Our approach hinges on identifying the restaurant the individual is at, and retrieving the menu of said restaurant. Although latitude and longitude can be reliably obtained with GPS sensors in mobile and wearable devices today, there might be times when the association between location data and the exact restaurant the person is visiting is erroneous (e.g. person is inside a shopping mall, or when two or three restaurants are in close proximity to each other). Although

this might seem like a limitation of our method, it is usually not of practical concern since restaurants that are physically close are typically significantly different in their offerings. Thus, it is often enough to identify the general physical area the individual is at (as opposed to the exact restaurant) and retrieve the menu of all restaurants and their respective food photos.

4.5.0.3 Semi-Automating Food Journaling

Dietary self-monitoring is effective when individuals are actively engaged and become aware of their eating behaviors. This, in turn, can lead to reflection and modifications in food habits. Our approach to food recognition is designed to facilitate dietary self-monitoring. Engagement is achieved by having individuals take a picture of their food; the tedious and time-consuming task of obtaining details about the food consumed is automated.

4.6 Conclusion

Although numerous solutions have been suggested for addressing the problem of poor adherence to nutrition journaling, a truly practical system for dietary self-monitoring remains an open research question. In this chapter, we present a method for automatically recognizing foods eaten in restaurants leveraging location sensor data and online databases.

The contributions of this work are (1) an automatic workflow where online resources are queried with contextual sensor data (e.g., location) to assist in the recognition of food in photographs.; (2) image classification using the SMO-MKL multi-class SVM classification framework with features extracted using color and point-based algorithms; (3) an in-the-wild evaluation of our approach with food images taken in 10 restaurants across 5 different types of food (American, Indian, Italian, Mexican and Thai); and (4) a comparative evaluation focused on the effect of location data in food recognition results.

CHAPTER V

LEVERAGING ENVIRONMENTAL CONTEXT

The previous chapters of this thesis explored the first two types of contexts: the cues derived from the data (spatio-temporal context) and the cues that are captured concurrently using external sensor devices (specifically, egocentric and geographic context) in order to effectively support dynamic scene understanding. In this chapter, we discuss our work on leveraging the third type of context – environmental context, for automated production of basketball highlights.

5.1 Introduction

The environment within which an activity is taking place may have third-party observers who are observing and reacting to the actors involved in the activity. The environment may also contain sensors that are capturing information about the ongoing activity and also recording the reactions of the third-party observers. In this chapter we show that several contextual cues can be derived from such observations within the environment and successfully leveraged to understand the dynamic scene that is taking place within the environment.

Sporting events are ideal for this study. The players are active within the environment and the audience reacts to their actions with a range of emotions ranging from excitement to frustration. Sensors such as cameras and microphones are setup by the broadcasters which capture the player activity and the audience reaction (both audio and video). There are also several “expert” third party observers within the environment such as referees, coaches, commentators, and on-court statisticians. The data from these observers coupled with the video data from the broadcast videos provides rich contextual cues that can be leveraged to understand the sporting scene.

An effective demonstration of understanding a dynamic scene such as sports is to automatically produce the highlights for the game. Generating the highlights for an entire sports game involves understanding the salient moments of the game, generating an excitement-based rank-ordering of the plays, segmenting and extracting them from the broadcast video, and selecting the top clips to generate the game highlights. Thus, in the context of this thesis, we define a sports highlight as a “highlight reel” that showcases the top n exciting moments of the game in a chronological order. The scope of this study is limited to basketball games. Basketball is the third most popular sport in the US (after Football and Baseball) [1] and is held in indoor stadiums and indoor gymnasiums in schools and colleges that provides a representative test-bed to develop our methodology and evaluate it.

In this study, contextual cues are derived from two sources within the basketball environment: (1) microphones that capture the audience and commentator audio, and (2) the play-by-play stats data from the on-court statisticians. From these two environmental sources, we extract four different cues: “Audio”, “Score Differential”, “Player Ranking” and “Basket Type”. Finally, a fifth cue “Motion” is extracted from the broadcast video which captures the magnitude of player and camera motion. For each basket within a given game, the data from these five cues is combined to generate an excitement score for the basket. Once all the baskets have been scored, we can then rank them by their excitement scores and pick the top n exciting clips and use them to generate the game highlights.

In order to conduct this study, we built a database of 25 NCAA games (played between February and March of 2015) totaling 35.44 hours of basketball footage along with the corresponding play-by-play stats data. There are a total of 1,173 baskets across these 25 games. We conducted extensive user-studies using Amazon’s Mechanical Turk in order to obtain ground-truth on the excitement levels for each of these 1,173 baskets. The ground-truth data was then used to study the effectiveness

of each of the cues as an indicator of how exciting a basket is. Finally, the five cues are combined by using a weighted sum wherein the weights are learned from the data using 25-fold cross-validation (where we train using 24 games and test on the held-out game and repeat). We conduct a second round of user-studies and show (1) the effectiveness of cue-combination over each of the individual cues, and (2) that the highlights that we generate with our cue-combination are comparable to the highlights produced by ESPN for those games.

Contributions: Our contributions are as follows: (1) We present a method to leverage environmental contextual cues to understand the excitement levels within a basketball game and automatically produce basketball highlights, (2) We introduce a new dataset of 25 NCAA games (35.44 hours of video with 1,173 baskets) along with the play-by-play stats and the ground-truth excitement data for each basket (we will make this dataset public to the research community), (3) We explore five different cues and study their effectiveness in determining the excitement of baskets through an extensive user study, and (4) We conduct user studies and show that the final highlights that we produce are comparable to the ones produced by ESPN.

5.2 *Related Work*

Sports analytics and summarization has been an active area of research for the past two decades. Most of the work has been on analyzing broadcast videos from sports such as soccer, basketball, hockey, football and tennis. Professional broadcast videos (such as videos from ESPN) contain replays which can be extracted by detecting the logo-sweeps (shown before and after the replays) and the “arousal” level of the replays can be computed using the audience’s audio energy and the amount of camera motion in order to rank the replays in terms of their excitement level [136]. Slow-motion replays can also be detected using Hidden Markov Models (HMMs) and Support Vector

Regressors [113] and summaries can be generated by concatenating the detected replays. When replays are not available in the broadcast video, baskets can be detected by detecting breaks in the game and using object detectors to detect the referee and the penalty boxes to make informed choices about the importance of different plays during the games [24]. However, these approaches are limiting since detecting replays and slow motions and using those clips in the highlights will give us a highlight reel that has only those baskets for which replays or slow motions were shown. There could be many other exciting baskets that are missed because the broadcast director chose not to show the replays or slow motions for those baskets.

Audio plays a crucial role in detecting highlights in sports. The energy of the crowd and the excitement in the commentator’s voice provides useful cues that can be used to pick exciting moments in the game. Audio-based architectures for sports summarization have been developed that extract audio features and classify the audio segments as applause, cheering, music, speech, etc. and also perform background noise modeling to further refine the results [130]. Along with audio, the amount of motion within the broadcast footage also helps identify exciting moments. The motion content of videos is encoded into the MPEG-7 motion activity descriptors. These motion vectors can be quantized and combined with the audio features to generate cumulative rankings of exciting moments [129, 71]. Audio and motion curves can also be combined to generate excitement time curves wherein the maximas represent the game highlights [38, 39] and the minimas around the maximas can be used to determine the segment boundaries of the highlight clips [79]. Motivated by these approaches, we investigate the use of audio and motion in our system.

An interesting area of research in sports summarization involves studying the problem from a affective rather than a cognitive point-of-view. The cognitive point-of-view is fact-based, wherein the features used for highlight detection are facts such as audio energy, amount of motion, position of ball, etc. In contrast, the affective

point-of-view is emotion-based and tries to understand the human emotions within the game. Affect has three underlying dimensions: valence (ranging from pleasant to unpleasant), arousal (ranging from excited to peaceful) and control (no-control to full-control). All of human emotions can be mapped in as a set of points in this 3D VAC space. Computational methods have been developed to compute the valence and arousal using video and audio features and using them to find highlights in both sports and movies, thereby generating summaries from an affective point-of-view [40].

In the past few years, with the proliferation of social media and blogging websites, researchers have turned their attention to “crowd-sourced” sports summarization techniques. People watching broadcast games use Twitter to tweet their reactions. Mining the Twitter data for relevant tweets and looking for times when there is a spike in the volume of tweets gives us the moments in time that the crowd deems to be interesting [85, 41]. Crowd-sourced summaries have several differences over traditional summaries generated by sports professionals. In crowd-sourced summaries the highlights that get selected include interesting plays that require high degree of skill (expected and easy plays are ignored), controversial plays and unusual occurrences (like fights and stunts) and “lowlights” which are moments in time when the fans are frustrated and angry at their favorite teams [108]. Other methods include analyzing web-casting text and social media blogs, aligning them with the broadcast videos and looking for highlights using player popularity and crowd sentiments [112, 131].

While most of the published works use cues derived only from the video data, the only other environmental contextual cue that is used is the commentator and audience audio. In our work, we look at the play-by-play stats that is obtained from the on-court statisticians, a source of data that has largely been ignored by the research community. We show that the play-by-play stats contain a wealth of information that can be leveraged to generate highlights that are comparable to the highlights by ESPN.

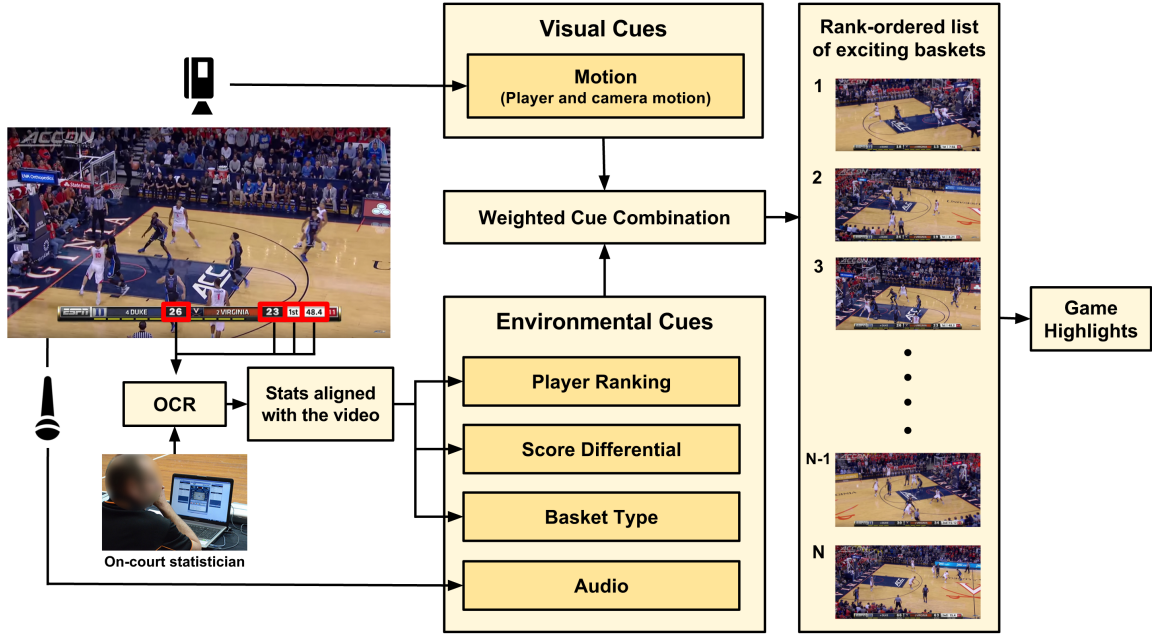


Figure 21: An overview of our system that uses visual and environmental contextual cues for automatically producing basketball highlights.

5.3 Methodology

An overview of our system is given in figure 21. The five different cues (four environmental and one visual) form the core component of our system. Let us look at each of these five cues in detail:

5.3.1 Cue 1: Audio

Gymnasiums and stadiums are equipped with microphones that capture the commentator and audience audio. Exciting baskets typically draw loud cheers from the audience and result in an elevation in the loudness and pitch in the commentator’s voice. The changes in their audio levels are important contextual cues that are indicative of how exciting a basket is [38, 39, 79, 130].

In our study, the audience and commentator audio is obtained from the broadcast video and thus unavailable on two separate channels. Let us denote this signal as a . Before we can compute statistics on a , it has to be pre-processed in order to obtain the

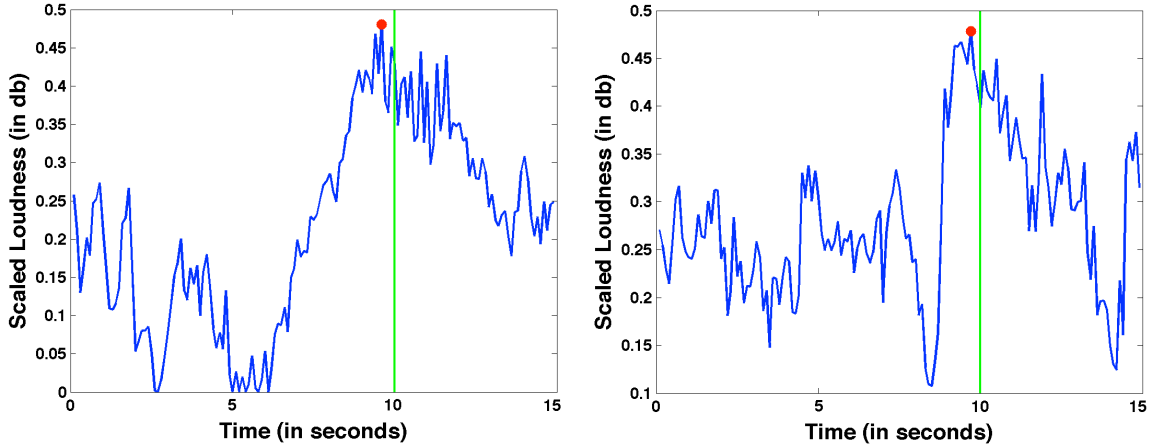


Figure 22: Audio loudness plots for two sample baskets. The red dot represents the time when the basket was scored and the green line (at the 10 second mark) represents the time when the scoreboard was updated to reflect the new scores. We can see that (1) the audio excitement peaks when the basket occurs, and (2) there is a slight delay of a few milliseconds between when the basket occurs and the scoreboard updates (the time difference between the red dot and the green line). The audio excitement drops soon after as the game continues.

true audio loudness, a_l , based on human perception of loudness. We perform this pre-processing by following the audio filtering guidelines provided by the International Telecommunications Union (ITU) [53]. The first stage of pre-processing applies a pre-filtering of the audio signal prior to the Leq(RLB) measure. The pre-filtering accounts for the acoustic effects of the head, where the head is modeled as a rigid sphere. The second stage of the algorithm applies the RLB weighting curve, which consists of a single high-pass filter. With the pre-filter and the RLB filtering applied, the mean square energy in the measurement interval T is then measure. Once the weighted mean square level has been computed for each channel, the final step is to sum the N audio channels. The audio loudness levels obtained using this approach is shown for two sample baskets in figure 22.

The true audio level, a_l , obtained using the ITU’s guidelines has been shown to be effective for use on audio programs *that are typical of broadcast content* which makes it the ideal audio pre-processing step for our application. Once the audio signal has

been pre-processed, the measure of excitement for a given basket b is computing as

$$A_b = \sum_{i=1}^m p_i(a_l) \tag{4}$$

where $p_i(a_l)$ is the i^{th} highest loudness peak in a 4 second window around the basket (3 seconds before the basket and 1 second after the basket). The overall audio loudness level for each basket, A_b , is obtained by summing the top m peaks. Empirically, for our NCAA dataset, m was determined to be 7. Finally, for each game, the A_b values for all the baskets are normalized between 0 and 1 by computing the *min* and *max* values across all baskets for that game.

5.3.2 OCR: Aligning the Stats With the Broadcast Videos

A main source of environmental context in our study is the play-by-play stats data that is generated by the on-court statisticians. The play-by-play data can be available in near-real-time or it can be available post-game. In either case, the stats need to be aligned with the broadcast video in order to determine when the particular play mentioned in the play-by-play stats actually occurred in the video. Unfortunately, the play-by-play stats are specific to a game and not specific to any particular broadcast video of the game. Hence they do not contain the video time-stamp of when the play occurred in the broadcast video.

In order to align the stats with the video, we introduce a novel Optical Character Recognition (OCR) based technique. The broadcast videos have a graphics overlay which contain four key pieces of information: (1) home team score, (2) visiting team score, (3) game period, and (4) game clock. An example is shown in Figure 23. Using the Tesseract OCR system [104], these four values are read for each frame of the video and are stored along with the corresponding video time-stamp. Next, we parse the play-by-play stats file and match the stored OCR info with each of the stats. This results in a mapping of the stats to the broadcast video. *Example:* Say for

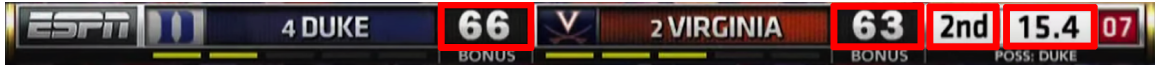


Figure 23: A typical graphics overlay shown in basketball broadcasts. They contain four key pieces of information (shown highlighted using red boxes): (1) home team score, (2) visiting team score, (3) game period, and (4) game clock.

a particular game, for a particular frame of the video, using OCR on the graphics overlay, we know that the home team score changed from “35” to “38” while the visiting team score was “29” during the “1st half” of the game at game clock “12:22” and when the video timestamp was “28:34”. While parsing the play-by-play stats, we see an entry “Player: Jahlil Okafor, Basket Type: 3-Pt Jump Shot, Game Period: 1st Half, Home Score: 38, Visiting Score: 29, Game Clock: 12:22”. By matching this entry with the OCR data, we can see that this particular 3-Pt Jump Shot basket by Jahlil Okafor took place at time-stamp “28:34” in the video. This allows us to align the rich contextual info from the stats with the corresponding basket within the video.

We extract three different cues from the stats data that are indicative of the excitement levels within the game: “Score Differential”, “Player Ranking” and “Basket Type” . Each of these cues are described below.

5.3.3 Cue 2: Player Ranking

Baskets by “star” (top-ranked) players tend to be generate more excitement among basketball fans than baskets by other lower ranked players. Also, our analysis of ESPN highlights of 10 NCAA games showed that ESPN tends to favor baskets by the star athletes and showcases them more in the highlights. The game stats may or may not contain the player ranking, but they almost always have the data on each player’s Points-Per-Game average (PPG). PPG has very strong correlation with player ranking and can be used instead when player ranking data is not available. For each game, we normalize each player’s PPG between the *min* and *max* PPG of all

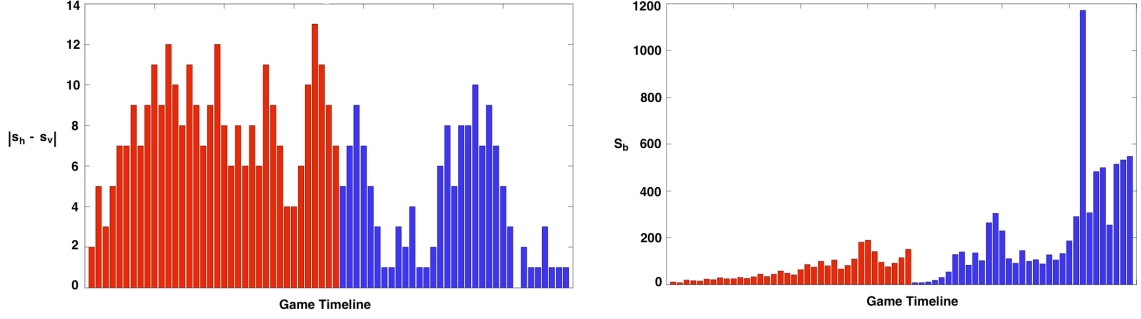


Figure 24: Score differential plots for a sample game. The x -axis represents the game timeline and the red and blue sections of the plots represent the 1st half and 2nd half of the game respectively. **Left:** $|s_h - s_v|$, the absolute score differential. **Right:** S_b , the inverted score differential weighted by the game-clock for each period of the game (see equation 5).

the players in that game (across both the teams). For each basket b , the scaled PPG value, P_b , of the player who made the basket gives us the “player ranking” excitement score for that basket.

5.3.4 Cue 3: Score Differential

People tend to find a game to be more exciting when the game is close (“neck-to-neck”) and less exciting when one team has a huge lead over the other. Furthermore, the game tends to be more exciting if the scores of the two teams are close towards the end of the game period. For a given basket b , if the home team score is s_h and the visiting team score is s_v , then the “score differential” excitement, S_b , for that basket is computed as

$$S_b = \frac{1}{(|s_h - s_v| + 1)} * (1200 - g_s) \quad (5)$$

where g_s is the game clock in seconds. As the score differential gets smaller, the excitement score S_b increases. Each game period is 20 minutes long and the game clock counts down from 20:00 (1200 seconds) to 00:00 (0 seconds). So, the score differential is weighed by the amount of time remaining in the game period. Lower score differentials towards the end of the game period will get higher weights, and

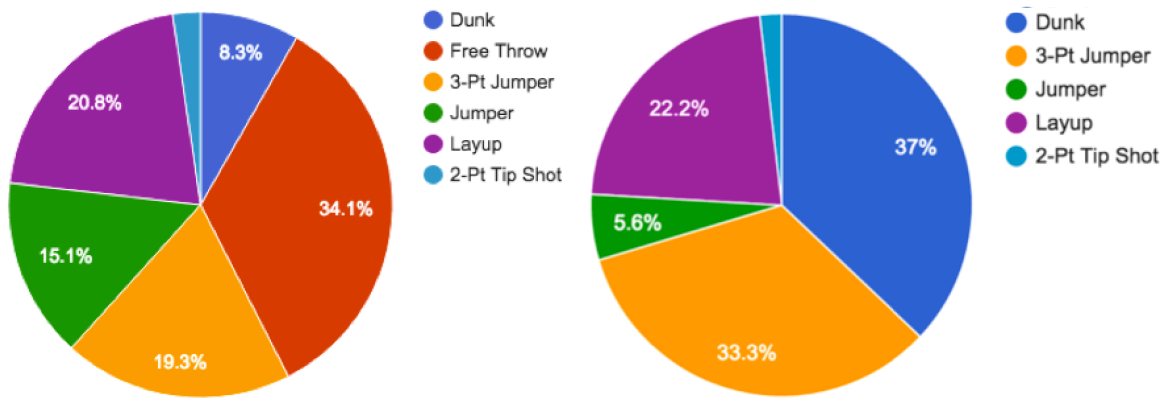


Figure 25: This figure shows the preference in baskets shown in the highlights generated by ESPN based on the basket type. **Left:** The distribution of baskets based on basket type across 10 full-length NCAA games. **Right:** The distribution of baskets based on basket type across the 10 highlights produced by ESPN for the same 10 NCAA games.

in-turn, higher excitement scores. The score differential plots for a sample game are shown in figure 24. The first half of the game is shown in red and the second half of the game is shown in blue. We can see the absolute score differential on the left and our “score differential” excitement, S_b , on the right. As with the other cues, the S_b scores are normalized between 0 and 1 for each game.

5.3.5 Cue 4: Basket Type

There are five types of baskets that are shown in basketball highlights: “Dunk”, “Jumper”, “Layup”, “Two Point Tip Shot”, and “Three Point Jumper” (“Free Throws” are typically not featured in the highlights). Each of these five baskets require different techniques and skills. Basketball fans tend to find some basket types more exciting than others. For example, the dunk shot is universally considered to be one of the most exciting basketball plays and is prominently featured in the highlights produced by ESPN. This is illustrated in figure 25. On the left, we can see the distribution of baskets based on the basket type across 10 full-length NCAA games. On the right, we can see the distribution of baskets based on the basket type across the 10 highlights produced by ESPN for the same 10 NCAA games. We can clearly see

that ESPN favors certain basket type over others. Although “Free Throw” occurs 34.1% of the time, they are almost never featured in the highlights due to the fact that a “Free Throw” is not very exciting to watch. However, “Dunk” occurs only 8.3% of the time, but is featured in 37% of the highlights. This is due to the fact that the viewers love watching a “Dunk” and consider it to be much more exciting than the other basket types.

These 5 baskets can be rank-ordered in $5! = 120$ different ways. Each of these 120 different basket rankings were evaluated on our NCAA dataset and the ranking that best matched the user-generated ground-truth was chosen. The ranking with the best match was: “Dunk” > “Two Point Tip Shot” > “Three Point Jumper” > “Layup” > “Jumper”. Using this ranking, for each basket b , the corresponding basket type’s rank position, B_b , gives us the “basket type” excitement score for that basket. The B_b scores are normalized between 0 and 1 for each game.

5.3.6 Cue 5: Motion

The amount of player motion during a given play is usually an indication of how exciting the play is. For example, a “Free Throw” which has minimal player motion is less exciting than a “Dunk” wherein all the players are in rapid motion. The amount of camera motion is also indicative of the excitement levels of the game. For example, a large panning motion is involved when something exciting happens, such as a player running from one end of the court to another with the ball. In contrast, a free-throw has almost no camera movement and is less exciting than the other types of plays.

For each basket, we computed the optical flow using KLT tracking [102] across all the frames. Camera motion was determined by computing the dominant optical flow and player motion was computed by subtracting the camera motion from the overall flow. For each basket b , the corresponding camera motion magnitude M_b^c ,

player motion magnitude M_b^p , and the overall motion M_b , gives us the “motion” excitement scores for that basket. Our experiments showed that M_b was a better indicator of excitement when compared against M_b^c and M_b^p individually. The M_b scores are normalized between 0 and 1 for each game.

5.3.7 Cue Combination

Once all the 5 cues have been extracted for all the baskets and have been normalized for each game so that they have the same scale, we can combine them using a weighted sum. The final score, C_b , for each basket b is given by

$$C_b = \omega_1 * A_b + \omega_2 * P_b + \omega_3 * S_b + \omega_4 * B_b + \omega_5 * M_b \quad (6)$$

where $\sum_{i=1}^5 w_i = 1$. For our 25 game NCAA dataset, the weights are learned using 25-fold cross-validation. One of the games is held out as test and the weights are learned using the ground-truth excitement data from the other 24 games. The process is repeated 25 times, each time holding out a different game for testing. The final cues weights are computed by averaging the weights across all the 25 runs.

5.3.8 Generating Highlights

With the final cue combination score for all the baskets of a game, we can rank-order them in terms of their excitement scores. The next step is to put them together to form the game highlights. The top n exciting baskets are selected from the rank-ordered list, extracted from the broadcast video, sorted by time-stamp (so that the baskets appear in order), and put together to form the game highlights. The value of n depends on the length of the desired summary.

Each of the n clips that were extracted from the broadcast video were 7 seconds long and had 1.5 second duration between when the basket occurred and the clip ended. These numbers were learned from the data by studying ESPN highlights. Figure 26 shows the histogram of durations of the baskets shown across 10 ESPN

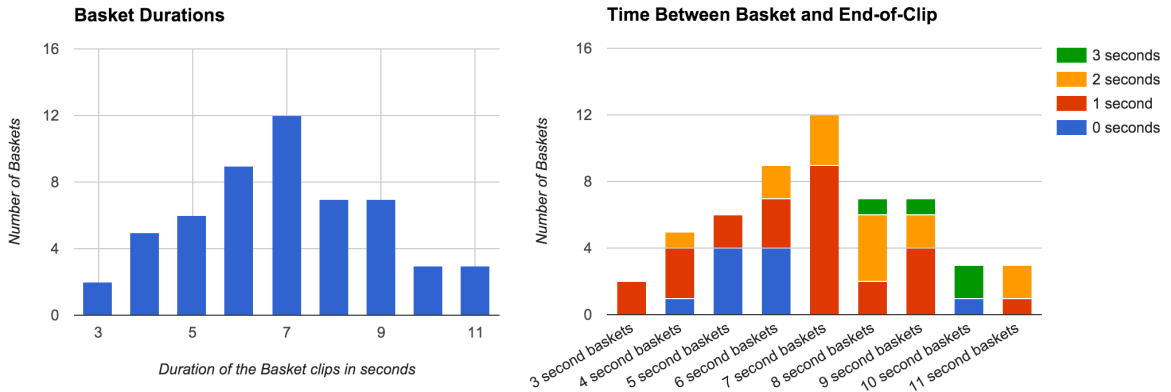


Figure 26: Left: Histogram of duration of the baskets shown across 10 ESPN highlights. We can see that ESPN prefers to show basket clips that are 6 to 7 seconds long. **Right:** Within each duration, we can see the time elapsed between when the basket occurred and the clip ended. For example, there are 12 baskets which were 7 seconds long. Out of these 12 baskets, 9 baskets had 1 second duration between when the basket happens and the clip ends and the other 3 baskets had 2 seconds duration.

highlights and the time elapsed between when the basket occurred and the clip ended.

We can see that ESPN prefers to have basket clips that are 6-7 seconds long with 1-2 seconds between the basket and the end of the clip.

5.4 Evaluation

In this section, we describing our NCAA dataset and the process by which we collected ground-truth pairwise excitement for each of the baskets. This is followed by the evaluation of each of our cues and a demonstration of the effectiveness of weighted cue combination as a predictor of excitement. Finally, we evaluate the highlights generated using cue combination against highlights generated using each individual cue and also compare against the highlights produced by ESPN.

5.4.1 Dataset Description

In order to gather ground-truth excitement data and evaluate our approach, we built a new basketball dataset. This dataset will be made public for the benefit of the research community.

We collected 25 full-length broadcast videos of NCAA games from 2015 (March

Madness) from YouTube. This is a total of 2,126.5 minutes (35.44 hours) of basketball videos. All the videos are 720p HD at 30fps.

For each of these games, we also collected the play-by-play stats data. Next, using the OCR technique described above, we aligned the play-by-play stats with the corresponding videos and extracted all the baskets such that each basket clip was 8 seconds long (5.5 seconds before the basket and 2.5 seconds after the basket). Free Throws were ignored since Free Throws are typically not shown in highlight reels. This gave us a total of 1,173 baskets across all the 25 NCAA games with the corresponding stats data for each basket. The stats data contains the following information: (1) player name, (2) basket type, (3) home team score, (4) visiting team score, (5) game clock, and (6) game period.

For 10 out of the 25 games, we also have the game highlights produced by ESPN. The videos were collected from YouTube and are 720p HD at 30fps.

5.4.2 Ground-Truth Pairwise Excitement


Getting ground-truth excitement data on all the 1,173 baskets in our dataset lets us analyze each of our five cues and evaluate the effectiveness of cue combination. However, collecting this ground-truth is non-trivial. Users find it hard to subjectively rank a bunch of clips based on how exciting the clips are. The more the number of clips, the harder the task becomes. However, users find it fairly easy to pick the exciting clip given only two choices. Thus, in order to gather the ground-truth data, we conducted A/B test user studies on Amazon’s Mechanical Turk where users were shown a basket from one of the games and another random basket from the same game and were asked the question “Which of these two clips is more exciting to watch?”. We took several steps to ensure the quality of these user studies:

- Each A/B pair was shown to 15 different users in order to get good data and reduce the likelihood of selection based on chance. This resulted in 17,595 A/B

Which of these two clips is more exciting to watch?

(Please watch with your audio turned on)

Clip A




6:29

4 DUKE 45 3 FLORIDA STATE 31 2nd 14:48

NCAAAM 125 21 Ohio St 52 15 Baylor 43

Clip B



6:35

4 DUKE 51 3 FLORIDA STATE 45 2nd 9:38

NFL In Inglewood. Moving any franchise to another city would require approval from

Please answer the following questions and press the submit button.

Q1. Which clip was more exciting to watch?

- Clip A
- Clip B

Q2. Are you a basketball fan?

- Yes
- No

Q3. Are you a fan (or alumnus) of one of the teams in the above clips?

- Yes
- No

Q4. How many basketball games did you watch last season?

- None
- 1 to 5
- 6 to 10
- 11 to 20
- 21 to 40
- Greater than 40

Q5. What's your age?

- 18 to 29
- 30 to 39
- 40 to 49
- 50 to 59
- 60 or older

Q6. Any other comments that you would like to make? Please enter here.

Figure 27: A sample A/B test page shown to the users.

tests (15 studies across 1,173 baskets).

- We required all Mechanical Turk users to have at least 95% approval rating and a minimum of at least 1,000 previously approved tasks. The A/B tests took an average of 1 minute and 29 seconds and the users were paid \$0.03 per study.
- The order in which the clips were shown in each A/B test was randomized. This further decreased the likelihood of user bias towards any one choice..
- The users were asked additional questions as a part of each A/B test in order for us to gain more insight into who our users were. These additional questions were:

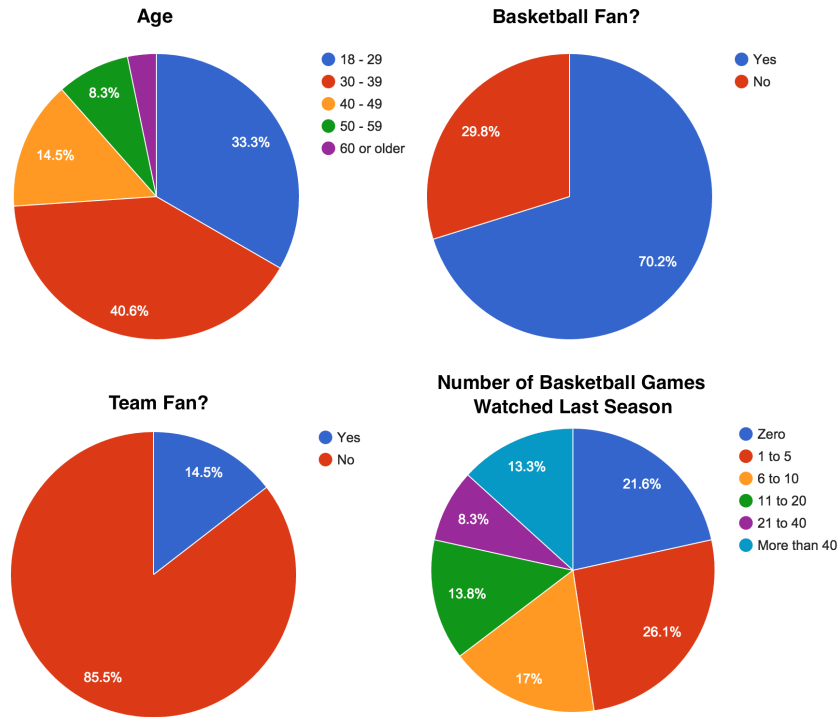


Figure 28: The distribution of users across the different A/B tests for capturing the ground-truth pairwise excitement levels of the baskets. Top Left: Distribution across age groups. Top Right: Distribution on whether basketball fan or not. Bottom-Left: Distribution on whether team fan or not. Bottom Right: Distribution based on the number of games watched last season.

- “Are you a basketball fan?” [options “Yes, “No”].
- “Are you a fan (or alumnus) of one of the teams in the clips?” [options “Yes, “No”].
- “How many basketball games did you watch last season?” [options: “None”, “1 to 5”, “6 to 10”, “11 to 20”, “21 to 40”, “Greater than 40”].
- “What’s your age?” [options: “18 to 29”, “30 to 39”, “40 to 49”, “50 to 59”, “60 or older”].

Figure 27 shows a sample A/B test page that was shown to the users. After all the A/B test user studies were completed, we analyzed the data and found that there were 399 unique users who had participated in our studies. The distribution of the users based on the questions we asked is as shown in Figure 28. We can see that the

Table 5: The inter-rater reliability metrics for our A/B tests on assessing the pairwise excitement of the baskets in our NCAA dataset.

Agreement between N or more users	Number of baskets	Average pairwise agreement percentage	Average pairwise Cohen’s kappa	Fleiss’ kappa	Interpretation
8	1173	54.26%	0.067	0.067	Slight agreement
9	880	56.79%	0.105	0.105	Slight agreement
10	625	60.14%	0.154	0.154	Slight agreement
11	384	65.01%	0.212	0.213	Fair agreement
12	203	71.18%	0.270	0.270	Fair agreement
13	92	77.76%	0.304	0.305	Fair agreement
14	18	88.15%	0.416	0.382	Fair agreement

majority of our users are between the age range of 18 to 39, are mostly basketball fans, and mostly not fans of any of the two teams shown in the clips.

5.4.3 Inter-Rater Reliability

In order to compute the overall consensus across our 15 A/B tests for each of the 1,173 baskets, we compute two inter-rater reliability metrics – the Fleiss’ kappa and the average pairwise Cohen’s kappa [30]. These statistical measures take into account the amount of agreement that could be expected to occur through chance. Table 5 shows the inter-rater reliability metrics for different values of N , where N is the number of users who agreed that one basket was more exciting than the other. For example, from Table 5 we can see that there are 384 baskets for which 11 or more users (out of the total 15) agreed that one basket was more exciting than the other in the randomized A/B tests. This has a Fleiss’ kappa of 0.213 which is interpreted as “Fair Agreement”.

5.4.4 Evaluating Individual Cues

The pairwise excitement ground-truth data allows us to study the effectiveness of each of our five cues in predicting how exciting a basket is. For our evaluations, we ignore the baskets which were hard to decide and focus only those baskets which had

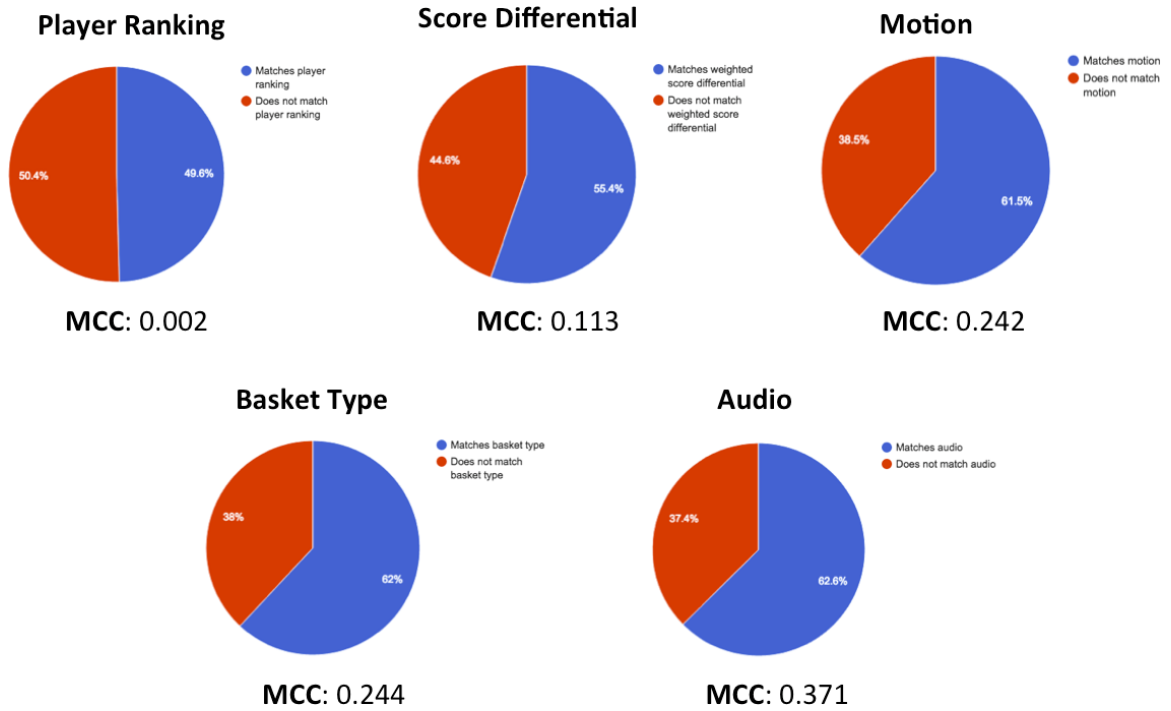


Figure 29: The performance of each cue as a predictor of the excitement levels of baskets.

at least 2/3rd agreement among the users, i.e. at least 10 out of 15 must agree that one basket is more exciting than the other. This lets us study the effectiveness of our cues in the absence of noise from the hard-to-decide baskets. From Table 5, we can see that this gives us 625 baskets for evaluations.

For this evaluation, each of the 625 A/B pairs that was shown to the users, is given as input to our system. For each individual cue, the system decides which basket is more exciting. The output of our system for each cue is then compared against the majority decision made by the users for that basket. If the system decision is same as the user decision, then we have a match. For each cue, we also compute the Matthews Correlation Coefficient (MCC) that gives the amount of agreement or disagreement between the system decision and the user decision [78, 92]. If MCC is -1, it means that the system decision and the user decision are in total disagreement. If MCC is +1, it means that the system decision and the user decision are in total agreement. If the MCC is 0, it means that the match is decision is no better than random.

Table 6: The percentage of baskets that match the user decision and the MCC score for each cue.

	Percentage of baskets that matches user decision	MCC
Player Ranking	49.6% (310 / 625)	0.002
Score Differential	55.4% (346 / 625)	0.113
Motion	61.6% (385 / 625)	0.242
Basket Type	62.2% (389 / 625)	0.244
Audio	62.7% (392 / 625)	0.371

The performance of each individual cue is shown in Table 6 and illustrated in figure 29. We can see that “Motion”, “Basket Type” and “Audio” are relatively strong indicators of how exciting a basket is while “Score Differential” and “Player Ranking” are very close to being no better than random. Out of the five cues, “Audio” is the strongest indicator with a MCC score of 0.371.

5.4.5 Evaluating Weighted Cue Combination

To learn the weights of the various cues, we perform 25-fold cross-validation where we hold out all the baskets from 1 game for testing while using all the baskets from the other 24 games for learning the weights. The process is repeated 25 times, with a different game being held out for testing in each run. In each run, all combinations of weights are tried and the combination that results in the most matches with the user decision on the held out test game is deemed as the winning set of weights for that run. After all the runs are complete, we average the weights across the 25 runs to get the final set of weights.

The average percentage of baskets that matched user decision across 25 runs was 75.33%. The lowest average percentage was 52.81% with the highest was 91.90%. When the weights are averaged across all the 25 runs and normalized to add up to one, as shown in figure 30, we see that **Player Ranking** gets **4.8%** of the total weight, while **Motion** gets **10.2%**, **Score Differential** gets **14.6%**, **Basket Type** gets **14.8%**, and **Audio** gets **55.6%** of the total weight respectively.

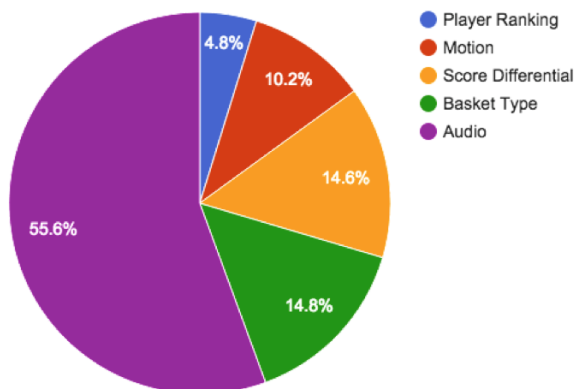


Figure 30: The percentage of the total weight that each cue gets during weighted cue combination (after running 25-fold cross-validation).

Table 7: McNemars tests on statistical significance between each individual cue vs. cue combination.

	χ^2	<i>p</i> -value
Player Ranking vs. Cue Combination	86.98	<0.0001
Score Differential vs. Cue Combination	63.00	<0.0001
Motion vs. Cue Combination	31.74	<0.0001
Basket Type vs. Cue Combination	57.34	<0.0001
Audio vs. Cue Combination	32.48	<0.0001

It is interesting to see that after combining all the cues, Audio makes up 55.6% of the total share. This shows that the audience cheers and the loudness and pitch of the commentator does indeed drive the excitement levels during a basketball game. It is also interesting to note that the top three dominant cues, Audio, Basket Type, and Score Differential, which make up 85% of the total weight are all contextual cues derived from the environment. This showcases the importance of leveraging context when applicable.

When the final average weights are used in cue combination and the system output is compared with the user decision, the percentage of baskets that match the user decision is **76.4%**. This is a significant improvement over the best percentage obtained by any single cue (62.7% with Audio, see Table 6). Furthermore, the MCC score goes up to **0.528** (from the previous best of 0.371 with only Audio). In order to

Table 8: A/B test results with **all users**: Cue combination highlights vs highlights generated using the individual cues.

Cue Combination Highlights vs.	Number of games for which cue combination highlights was selected by user majority	Median user agreement percentage
Player Ranking Highlights	22 / 25	61.29%
Score Differential Highlights	17 / 25	54.84%
Motion Highlights	17 / 25	58.06%
Basket Type Highlights	16 / 25	58.06%
Audio Highlights	16 / 25	61.29%

ensure that the improvement we see from cue combination over each of the individual cues is statistically significant, we ran the McNemar’s chi-square test (with Yates’ continuity correction). The null hypothesis says that the improvements we see after cue combination are due to chance., However as shown in Table 7, the χ^2 values are greater than the critical value (at 95% significance level) of 3.84 and the p -values are less than the significance level (α) of 0.05. Thus the null hypothesis can be rejected. The improvements seen after cue combination are *not* due to chance.

5.4.6 Evaluating Highlights

With the final weights for each cue, we can now rank order all the baskets in a game by their excitement score, pick the top N baskets, order them by the game clock (so that they are in order), and generate the highlight video. For the 10 games for which we have ESPN highlights to compare against, we pick N as the number of baskets that ESPN put in their highlights for each of those 10 games. For the other 15 games, we pick N to be 10 (the average number of baskets that ESPN shows in their highlights).

Cue Combination vs. Individual Cues: We generated highlights for all the 25 games using both cue combination and also using each of the five individual cues. This gave us 6 highlight videos for each game. Figure 31 shows 4 sample frames from the

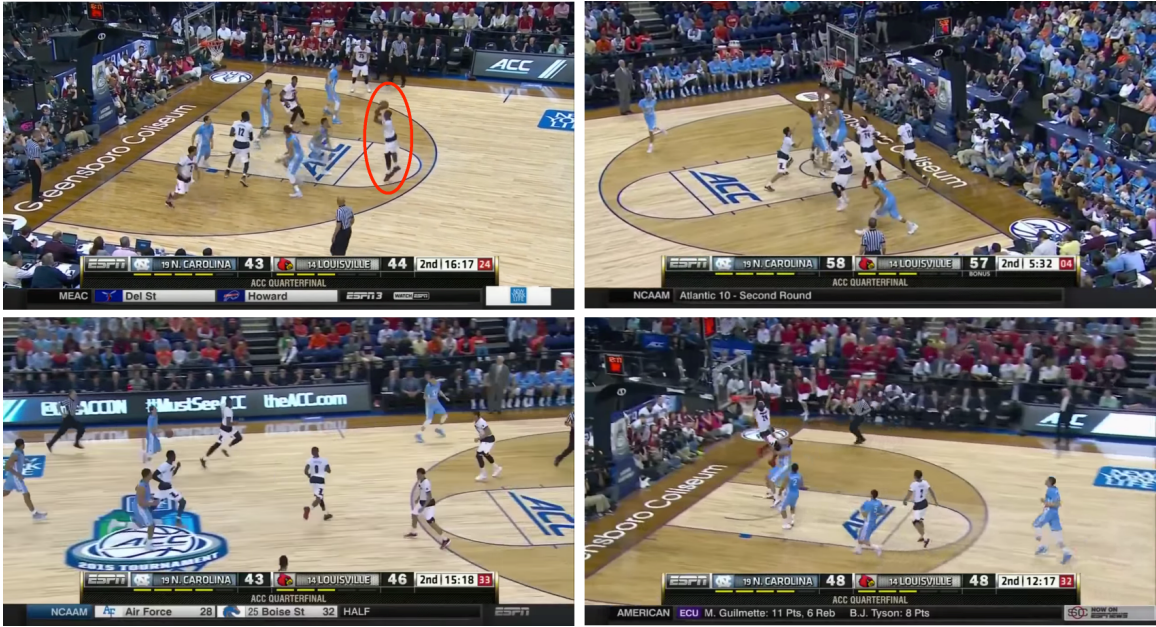


Figure 31: Sample frames from highlights generated using individual cues: Player Ranking (Top-Left), Score Differential (Top-Right), Motion (Bottom-Left), and Basket Type (Bottom-Right).

highlights generated for the Louisville vs. North Carolina NCAA game. Highlights generated using only Player Ranking mostly featured baskets by Terry Rozier (circled in red) who was the star player for the Louisville Cardinals (now a NBA draft pick for the Boston Celtics). The highlights generated using only Score Differential featured “neck-to-neck” baskets (for example, a score of 58-57 can be seen in the figure) while the highlights generated using only Motion featured baskets with lots of player and camera motion (as seen in the figure). Finally, the highlights generated using only Basket Type featured mostly Dunk shots (a sample shot is shown in the figure).

In order to see if users prefer cue combination highlights over the highlights generated by individual cues, we ran another A/B test user study. Given the harder nature of the study where users have to watch two minute-long clips, we doubled the number of users per test from 15 to 31. For each game, the users were asked to pick the highlight that they preferred. Similar to the previous study, sufficient care was taken to randomize the A/B pairs and the users were asked to fill in a similar

Table 9: A/B test results with **basketball fans**: Cue combination highlights vs highlights generated using the individual cues.

Cue Combination Highlights vs.	Number of games for which cue combination highlights was selected by user majority	Median user agreement percentage
Player Ranking Highlights	20 / 25	60.00%
Score Differential Highlights	14 / 25	60.00%
Motion Highlights	18 / 25	60.00%
Basket Type Highlights	16 / 25	60.00%
Audio Highlights	15 / 25	60.00%

questionnaire mentioning their age, if they are a basketball fan, if they are a fan of one of the teams, and the number of games they watched last season.

This study had 335 unique users out of which 245 are basketball fans. The results of the study with all 335 users are shown in Table 8 and the results with only the 245 basketball fans is shown in Table 9. We can see that highlights generated using cue combination were preferred over the highlights generated using the individual cues by both regular users and basketball fans. However, it is interesting to note that basketball fans seem to prefer score differential highlights slightly more than regular users. This could be due to the fact that basketball fans watch the game more closely and pay more attention to the scores shown on the graphics overlay.

Cue Combination vs. ESPN: Our dataset contains the highlights produced by ESPN for 10 out of the 25 games. We ran a similar A/B test study with 31 users where users were shown the ESPN highlight and our cue combination highlight and were asked to pick the highlight that they preferred. To make the comparison fair, we regenerated the ESPN highlights using the same video production pipeline that we used to produce our highlights. This ensured that both the highlights shown in the A/B test are visually similar.

The results of the A/B tests showed that among all users, our highlights were preferred in 5/10 games and the ESPN highlights were preferred in the other **5/10**

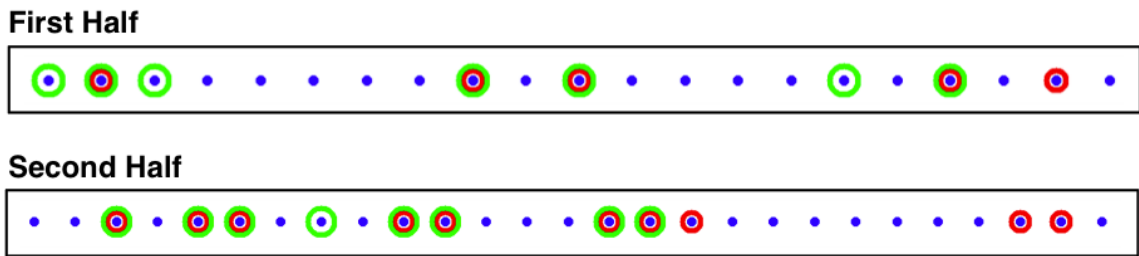


Figure 32: Basket picks for a sample game (Duke vs. Florida State, 9th Feb 2015). Each blue dot represents a basket that occurred during the game play. Baskets with red circles were picked by ESPN for their highlights and baskets with green circles were picked by our method for our highlights. We can see that 11 out of the 15 baskets were commonly picked (4 overlaps in the first half and 7 overlaps in the second half).

games. The median agreement percentage was 51.61%. Among basketball fans, our highlights were preferred in **7/10** games while the ESPN highlights were preferred in 3/10 games. The median agreement percentage was 53.33%. Although basketball fans showed a slight preference to our highlights, the median agreement percentage shows that the decision was really hard to make. This shows that the users had a tough time picking between our highlights and ESPN highlights which in-turn indicates that we are performing as well as ESPN in producing basketball highlights.

Comparing the individual baskets that were picked for the highlights across all the 10 games, we noticed 67.4% overlap in the baskets that we picked and the baskets that ESPN picked. This is illustrated in figure 32 where we show the basket picks for a sample game (Duke vs. Florida State, 9th Feb 2015). We can see that across both the game periods, out of the 15 baskets, 11 baskets were commonly picked by our cue-combination approach and by ESPN. The probability of these 11 baskets being picked by chance is 0.00005.

Another factor to consider is the distribution of the baskets shown in the highlights across the two halves of the game. As shown in figure 33, for the 10 games, our cue combination picks 48% of the baskets from the first half of the game and 52% of the baskets from the second half of the game. Looking at the ESPN highlights for those

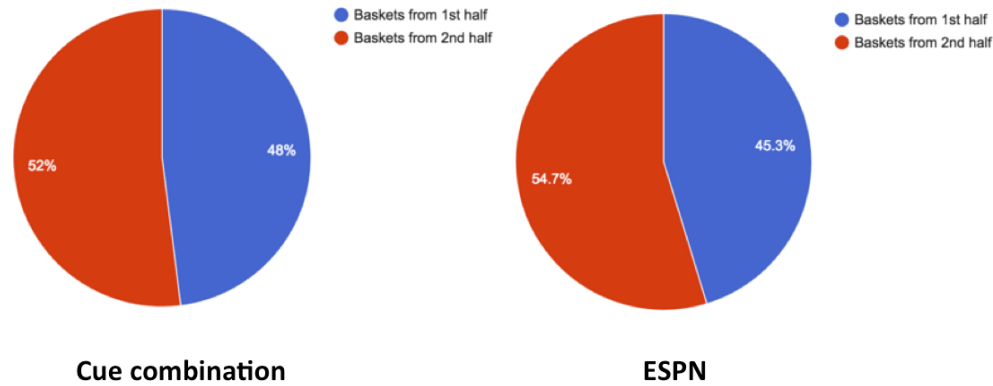


Figure 33: The distribution of baskets across the two halves of the game for 10 games. **Left:** The basket distribution for our cue combination highlights. **Right:** The basket distribution for ESPN highlights. We can see that our approach isn't biased towards any particular half of the game and closely follows ESPN's distribution across the two halves.

10 games, we can see that ESPN has a very similar distribution. They pick 45.3% of the baskets from the first half and 54.7% of the games from the second half. This shows that our approach isn't biased towards baskets from any single period of the game and closely follows ESPN's distribution.

Figure 32 and figure 33 further highlights the practicality of our approach and the similarity of our highlights to the ESPN highlights. This gives us an insight as to why the users had a hard time deciding between our highlights and ESPN highlights.

5.4.7 Evaluation Strategy

We have used different evaluation strategies in the first three chapters of this thesis depending on the application and the dataset, namely supervised, unsupervised and semi-supervised learning based evaluations, localization-based evaluations and visualizations. In this chapter, the subjective nature of evaluating the excitement levels of baskets and the highlight videos led us to follow a different evaluation strategy of conducting A/B test user studies wherein we rely on the consensus obtained from multiple users per A/B test. The A/B tests helped us gather ground-truth excitement data for all the 1,173 baskets across the 25 NCAA games and was valuable in

determining the effectiveness of cue combination against each individual cue and also against the highlights generated by ESPN.

5.5 Discussion

Our comprehensive user studies and analysis have shown the effectiveness of leveraging contextual cues derived from the environment in determining the exciting baskets of a game and automatically producing the game highlights.

However, there are some limitations of our work. The ground-truth data that we used for evaluations is pairwise excitement gather using our A/B tests. Pairwise excitement scores are inherently different from true excitement scores wherein the users look at all the baskets in a game and puts them in a rank-order from least exciting to most exciting. While this kind of data would be more useful than pairwise excitement scores, it is practically impossible to get the true excitement scores through user studies. A typical basketball game has 50 or more baskets and it is very hard for humans to look at so many baskets, remember their excitement levels, and rank order them. Another limitation is that our method does not adaptively change the weights of the cues based on changes in the environment. For example, if the audience start leaving after the first half, we should be able to detect this and adaptively decrease the weight given to the audio cue. This is a good direction for our future work.

While running the A/B tests where users are shown two basket clips and asked to pick the one that is more exciting, we wondered how the results would change if users were shown the same A/B pair, but with only the audio and only the video. To answer this question, we picked the top 203 baskets that had 80% (12/15) agreement between the users (i.e. baskets for which most users had a strong consensus in deciding between choice A and choice B), and generated two more A/B tests with the same pairs as before. But this time one of the A/B tests had just the audio of the game and the other A/B test had clips with just the video (with the audio stripped out).

We ran these new A/B tests for the 203 baskets with 15 users per test (as before) and analyzed the results. It is interesting to note that for the audio-only case, the agreement fell from 80% to **59.70%** and the Fleiss' kappa value dropped from 0.270 (fair agreement) to **0.177** (slight agreement), and for the video-only case, the average pairwise agreement fell to **59.82%** and the Fleiss' kappa value dropped to **0.116** (slight agreement). This shows that the strong consensus between the users broke down when they had to pick between clips that had either only the audio or only the video. This further proves the point that the contextual cues provided by audio (and other sources) can be a valuable tool in developing practical Computer Vision applications.

An interesting question is “how would a highlight generated using randomly selected baskets compare against our cue combination highlights and against ESPN highlights?”. To answer this question, we generated 10 highlights using random baskets selected from 10 of the games (and placed in chronological order) and ran additional A/B tests for evaluating these random highlights against our cue combination highlights and also against the ESPN highlights. As before, we ran the tests with 31 users per A/B test. The results showed that among all users our cue-combination highlights were preferred over random highlights in **7/10** games with a median user agreement of **61.29%**. Basketball fans preferred our highlights in **8/10** games with a median user agreement of **53.33%**. The A/B test results against ESPN highlights showed that among all users ESPN highlights were preferred over random highlights in **7/10** games with a median user agreement of **54.84%**. Basketball fans preferred ESPN highlights in **9/10** games with a median user agreement of **60.00%**. These results tell us that while users prefer our highlights and ESPN highlights over random highlights, the random highlights still do get picked in some of the games. This further highlights the subjective and complex nature of the problem domain. However, it also tells us that there is probably no uncanny valley for basketball highlights and

that users are much more forgiving of mistakes that our approach may make.

5.6 Conclusion

In this chapter, we explored the use of environmental context and presented a practical application wherein environmental contextual cues can be leveraged to automatically produce basketball highlights. We explored 5 cues that are indicative of excitement levels in basketball games. Four of these cues are derived from sources within the environment while the fifth cue is extracted from the video data. We introduced a new dataset of 25 NCAA games with 1,173 baskets with ground-truth pair-wise excitement scores for evaluating our approach with. We conducted comprehensive user studies with multiple participants which showed the effectiveness of our cues and our cue combination method that can produce highlights that are comparable to those produced by ESPN. Interesting directions for future work include exploring methods to collect more ground-truth excitement data in a way that maximizes inter-rater reliability, exploring more cues that are indicative of excitement levels, and dynamically adapting the weights of the cues as the game proceeds based on the changes in the environment.

CHAPTER VI

CONCLUSION

In this thesis, we explored several methodologies of leveraging contextual cues for dynamic scene understanding and presented solutions to several real-world Computer Vision problems that would have either been intractable or impractical without the aid of context.

In particular, we categorized context into three broad classes: (1) context that is part of the video data, (2) context that is concurrently collected from an external sensor co-located with the camera, and (3) context provided by the environment within which the scene is unfolding. Our hypothesis was that contextual information in either of these three classes can be effectively leveraged, along with the video data, to improve dynamic scene understanding. Through our various studies that explored each of these three contextual classes across several different application domains, we showed that our hypothesis does indeed hold true. This thesis highlights the valuable nature of context and how it can be brought to bear on hard to solve Computer Vision problems that have practical real-world applications. Figure 34 provides a summary of the work done and the contributions of this thesis.

In the first part of this thesis, we studied context that is part of the video data. We presented a significant extension to BoW-based activity recognition approaches, where we augmented classical BoW models with temporal, local and global structural information, using three different temporal encoding techniques using n -grams and randomly-sampled regular expressions. Along with activity recognition, we showed that our approach can also be used to detect anomalies, predict skill levels, and categorize objects based on their function. Our Augmented-BoW technique has also

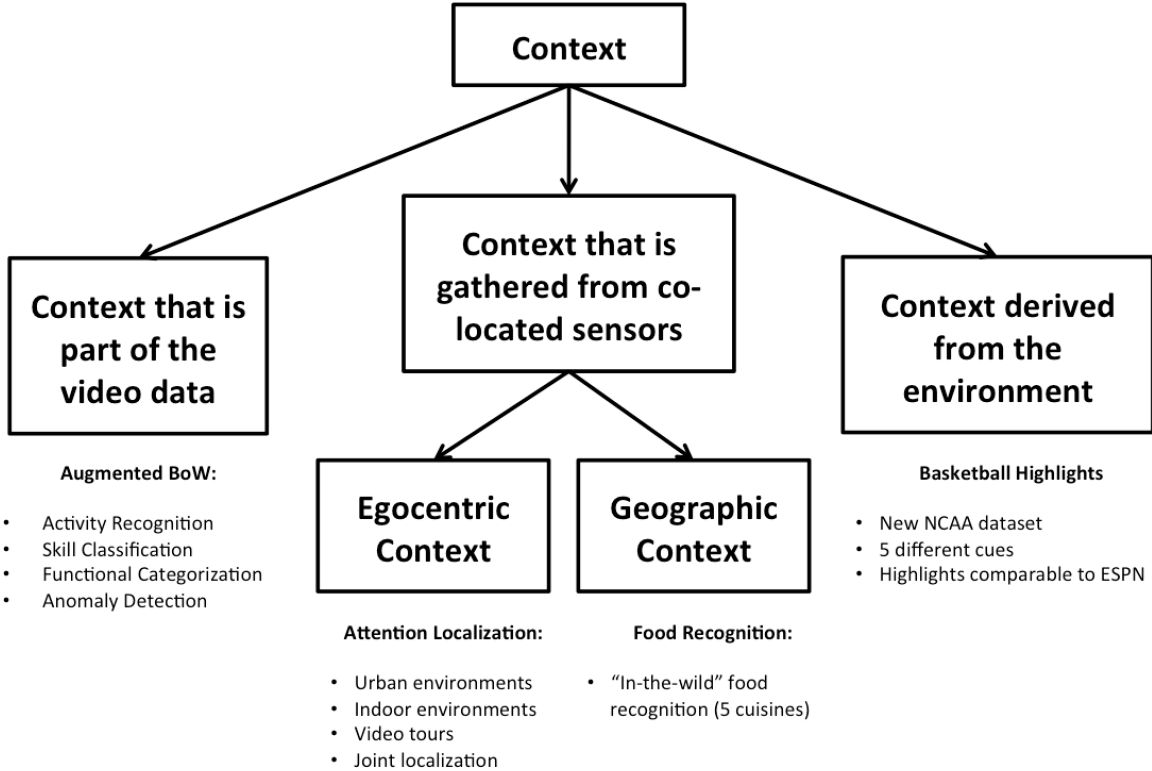


Figure 34: Summary of the contributions of this thesis.

been used to analyze non-video data as well, such as recognizing home-based activities using water-pressure data [109] and detect insider threats in large corporate systems using system usage logs [100].

In the second part of the thesis, we studied context that is concurrently collected from an external sensor co-located with the camera. In particular, the focus of our study was on two types of sensor data: (1) egocentric head orientation information, and (2) GPS. Using egocentric sensor data, we demonstrated a working system that can localize a person’s FOV, determine the point-of-interest, and map shifts in attention over time. We also showed how egocentric context from multiple sensors can be used to localize the joint attention of several people in both indoor and outdoor environments. For the second half of this study, we showed how GPS can be used to automatically recognize foods eaten in restaurants and our experiments on 5 different cuisines highlighted the potential practicality of our approach. Apart from GPS and

egocentric orientation, there are other types of sensor data that can be captured along with the video data. Examples include accelerometer data, magnetic field readings, infra-red heat signals, etc. As shown by this thesis, depending on the application in hand, using the relevant contextual data will typically result in a more practical, reliable, and scalable approach than using just the video data.

In the final part of this thesis, we studied context that is provided by the environment within which the activity is taking place. Using the example of sports, we showed how contextual cues derived from the on-court statisticians and the audience and commentator audio can be used along with the video data to understand the excitement levels of the game and automatically produce the game highlights. We conducted comprehensive user studies to generate pairwise excitement ground-truth for our NCAA dataset and to show the effectiveness of the environmental contextual cues.

Limitations: Leveraging context makes hard problems more tractable and is instrumental in building practical real-world applications. While each of the different types of context have been thoroughly discussed in this thesis, they do have certain limitations.

Let us first look at the context that is part of the video data. Our proposed Augmented Bag-of-Words model is very good at recognizing long term activities, i.e. activities that span several minutes, hours or even days. Long term activities can intuitively be represented as sequences of (short term) events and are rich in temporal and structural information which is exploited by our approach to build better bag-of-words (BoW) models. To highlight the length of the activities that we apply our techniques on, let us look at the four datasets used: (1) Ocean City surveillance: Each vehicle activity is typically 1 to 5 minutes long; (2) Surgical skill assessment: Each suturing activity is typically 15 to 20 minutes long; (3) Player activities from

soccer videos: Each game play is 45 minutes long; (4) Wide Area Airborne Surveillance (WAAS): The human activities span a duration of 46.5 hours (almost 2 days). In contrast, the state-of-the-art human activity datasets like Hollywood2, YouTube, UCF Sports and others contain short term activities like pick-up-phone, open-car-door, dive, swing, handshake, etc. These activities are typically 5 to 20 seconds long. Our approach, which is designed to make inferences from long term activity data, is not the right tool for such short term recognition tasks. The reason for this is as follows: The standard approach in human activity recognition is to cluster features using k-means with $k=4000$ to build BoW codebooks (in the literature, the value $k=4000$ has been shown to give the best empirical results). When short term videos of length 5 to 20 seconds are represented by words drawn from a large vocabulary of size 4000, the words get densely packed together and the time elapsed between the individual words tends to become zero. Thus our approach to discover the underlying temporal information is no longer applicable. Also there is insufficient information to generate random sampled regular expressions that are useful for discovering global patterns. However, it is important to note that the short term activities like open-car-door, pick-up-phone, etc are exactly the type of events that long term activities are composed of. For example, a long-term-activity such as: (event1:open-car-door, event2:kiss, event3:pick-up-phone, ...), could be given as input to our model which can then extract spatio-temporal contextual information for long term activity recognition, skill assessment, functional categorization and anomaly detection.

Next, let us look at the context that is gathered from co-located sensors. In the case of egocentric context, we make the assumption that a reference dataset is always available. Similarly, in the case of geographic context, we again make a similar assumption that the menu for each restaurant is available and that it is possible to automatically get representative canonical images (using web-based image search) for the food items on the menu to train our classifiers. While such assumptions on the

availability of reference data may not hold in some real-world applications, there is sufficient evidence to believe that these limitations may disappear with the current trend towards collecting more data and making it easily accessible to all. Indoor and outdoor spaces are being mapped, modeled and photographed more frequently with better cameras. Also, several companies are building centralized databases of restaurant menus while people are taking photographs of food much more frequently and tagging them appropriately. This explosion in data is extremely beneficial to systems that rely on contextual information since they now have access to better quality and more reliable contextual information.

Another limitation is the reliability of the sensor data. Gyroscopes and accelerometers drift over time, magnetometers are influenced by surrounding objects and structures and GPS is only accurate to a certain range. Thus relying heavily on the sensor data may lead to unreliable results. Finding the right balance between computationally intensive vision techniques and cheap contextual data can be tricky and needs to be fine-tuned based on the specifications and quality of the sensors that are available and the application that they are being used for. However, with the innovation in sensor technology, with time this limitation is bound to get addressed. Similar questions of reliability apply to environmental context as well. For example, to generate the basketball highlights, we rely heavily on the play-by-play stats data. This data is human generated by the on-court statisticians and is prone to human errors. Errors in the stats can propagate throughout the system and can lead to less than optimal highlight generation. This problem can be solved by capturing the environmental contextual data using two or more sources and cross-referencing and cross-checking the data to ensure its accuracy.

Future Work: We hope this thesis has been effective in convincing the reader regarding the importance of context. Our human perception system extensively leverages context by looking at the global organizational structure of the scene before

making local decisions. Building similar vision systems that leverage context gets us closer to solving hard problems that seem intractable when only the visual signal is considered.

While we have categorized and studied different types of context, this thesis does not attempt to present a unifying framework. This is a very interesting and useful direction for future work in this area. Providing an unifying framework involves the following key areas of research:

1. How best can we equip video and image capture devices with co-located sensors?

Google Glass is a good example of innovative research in this area. Glass is a small wearable device, yet contains sophisticated sensor-fusion capabilities due to on-device custom-designed sensors such as gyroscope, accelerometer and magnetometer. Another example is the Contour wearable camera that comes with GPS and also measures the speed and elevation of the user. Such innovative devices allow for the collection of relevant and useful contextual information.

2. How best can we store the contextual information with the media data? As the

contextual data (for example GPS) is being captured along with the video (or the image), it needs to be stored along with the media data. For example, for images, the contextual information can be stored as metadata information in the EXIF data of the images and videos. However the video capture and the sensor data capture could be at different frame rates and need to be synchronized appropriately to ensure that the resulting metadata is valid and useful.

3. How best can we convert the contextual metadata captured by various devices

to a standardized format? Different device manufacturers capture and store the contextual information in their proprietary formats. However, a universal standardized format is needed that can be used to convert the metadata to and from the manufacturer's format. The standard format needs to include all the

relevant data and also be flexible enough to be extensible for future needs that may arise.

4. How best can the standardized contextual data be used by vision systems for real-world applications? Having the contextual data in a standardized format will allow researchers to develop vision systems that can interact with other vision systems and pass data to each other efficiently. This can potentially lead to a modular model where the contextual information can flow through the various components of the system and can be leveraged as and when needed by the different components.

Research and engineering in each of the above areas will go a long way in having a unified and standardized contextual framework that vision systems can closely integrate and interact with. This will take us closer towards our endeavor of building practical and real-world computer vision systems that are faster, more accurate and more reliable.

REFERENCES

- [1] “Gallup sports popularity survey.” <http://www.gallup.com/poll/4735/sports.aspx>.
- [2] “Google Art Project.” <http://www.google.com/culturalinstitute/project/art-project>.
- [3] AALTO, L., GÖTHLIN, N., KORHONEN, J., and OJALA, T., “Bluetooth and WAP push based location-aware mobile advertising system,” in *Int. Conf. Mobile systems, applications, and services*, pp. 49–58, ACM, 2004.
- [4] AMFT, O. and TRÖSTER, G., “Recognition of dietary activity events using on-body sensors,” *Artificial Intelligence in Medicine*, vol. 42, pp. 121–136, Feb. 2008.
- [5] ARAB, L., ESTRIN, D., KIM, D. H., BURKE, J., and GOLDMAN, J., “Feasibility testing of an automated image-capture method to aid dietary recall,” *European Journal of Clinical Nutrition*, vol. 65, pp. 1156–1162, May 2011.
- [6] ARBELAEZ, P., MAIRE, M., FOWLKES, C., and MALIK, J., “Contour detection and hierarchical image segmentation,” *PAMI*, vol. 33, no. 5, pp. 898–916, 2011.
- [7] BACH, F., LANCKRIET, G., and JORDAN, M., “Multiple kernel learning, conic duality, and the smo algorithm,” in *ICML*, 2004.
- [8] BAHL, P. and PADMANABHAN, V. N., “Radar: An in-building RF-based user location and tracking system,” in *INFOCOM*, vol. 2, pp. 775–784, IEEE, 2000.
- [9] BETTADAPURA, V., ESSA, I., and PANTOFARU, C., “Egocentric field-of-view localization using first-person point-of-view devices,” in *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, January 2015.
- [10] BETTADAPURA, V., SCHINDLER, G., PLOETZ, T., and ESSA, I., “Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE Computer Society, June 2013.
- [11] BETTADAPURA, V., THOMAZ, E., PARNAMI, A., ABOWD, G., and ESSA, I., “Leveraging context to support automated food recognition in restaurants,” in *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, January 2015.

- [12] BIEDERMAN, I., “On the semantics of a glance at a scene,” *Perceptual organization*, 1981.
- [13] BLEI, D., NG, A., and JORDAN, M., “Latent dirichlet allocation,” *JMLR*, 2003.
- [14] BREIMAN, L., “Bagging predictors,” *Machine Learning*, vol. 24, pp. 123–140, 1996.
- [15] BRUCE, N. and TSOTSOS, J., “Saliency based on information maximization,” *NIPS*, vol. 18, p. 155, 2006.
- [16] BURKE, L. E., WANG, J., and SEVICK, M. A., “Self-Monitoring in Weight Loss: A Systematic Review of the Literature,” *YJADA*, vol. 111, pp. 92–102, Jan. 2011.
- [17] CERF, M., HAREL, J., EINHÄUSER, W., and KOCH, C., “Predicting human gaze using low-level saliency combined with face detection,” in *NIPS*, 2007.
- [18] CHANG, C.-C. and LIN, C.-J., “LIBSVM: A library for support vector machines,” *ACM Trans. on Intelligent Systems and Technology*, 2011.
- [19] CHEN, M., DHINGRA, K., WU, W., YANG, L., SUKTHANKAR, R., and YANG, J., “Pfid: Pittsburgh fast-food image dataset,” in *ICIP*, 2009.
- [20] CHEN, M., MUMMERT, L., PILLAI, P., HAUPTMANN, A., and SUKTHANKAR, R., “Exploiting multi-level parallelism for low-latency activity recognition in streaming video,” in *ACM SIGMM Conf. on Multimedia Systems*, 2010.
- [21] CHEN, M.-Y., MUMMERT, L., PILLAI, P., HAUPTMANN, A., and SUKTHANKAR, R., “Exploiting multi-level parallelism for low-latency activity recognition in streaming video,” in *First annual ACM SIGMM conference on Multimedia systems*, pp. 1–12, ACM, 2010.
- [22] CHENG, J., ZHOU, B., KUNZE, K., RHEINLÄNDER, C. C., WILLE, S., WEHN, N., WEPPNER, J., and LUKOWICZ, P., “Activity recognition and nutrition monitoring in every day situations with a textile capacitive neckband,” in *the 2013 ACM conference*, (New York, New York, USA), p. 155, ACM Press, 2013.
- [23] DUCHENNE, O., J., I. L., SIVIC, BACH, F., and PONCE, J., “Automatic annotation of human actions in video,” in *ICCV*, pp. 1491–1498, 2009.
- [24] EKIN, A., TEKALP, A. M., and MEHROTRA, R., “Automatic soccer video analysis and summarization,” *IEEE Transactions on Image Processing*, vol. 12, no. 7, pp. 796–807, 2003.
- [25] EVERINGHAM, M., VAN GOOL, L., WILLIAMS, C., WINN, J., and ZISSERMAN, A., “The pascal visual object classes (voc) challenge,” *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

- [26] FATHI, A., HODGINS, J. K., and REHG, J. M., “Social interactions: A first-person perspective,” in *CVPR*, pp. 1226–1233, 2012.
- [27] FATHI, A., LI, Y., and REHG, J. M., “Learning to recognize daily actions using gaze,” in *ECCV*, pp. 314–327, 2012.
- [28] FINK, G. A., *Markov Models for Pattern Recognition, From Theory to Applications*. Springer, 2008.
- [29] FISCHLER, M. A. and BOLLES, R. C., “Random sample consensus,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [30] FLEISS, J. L. and COHEN, J., “The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability.,” *Educational and psychological measurement*, 1973.
- [31] FORSYTH, D. A., MUNDY, J. L., ZISSERMAN, A., and ROTHWELL, C., “Using global consistency to recognise euclidean objects with an uncalibrated camera,” in *CVPR*, pp. 502–507, IEEE, 1994.
- [32] GAIDON, A., HARCHAOUI, Z., and SCHMID, C., “Action sequence models for efficient action detection,” in *CVPR*, 2011.
- [33] GALLUP, “Fast food still major part of u.s. diet,” Aug. 2013.
- [34] GORIS, A., WESTERTERP-PLANTENGA, M., and WESTERTERP, K., “Under-eating and underreporting of habitual food intake in obese men: selective underreporting of fat intake,” *The American journal of clinical nutrition*, vol. 71, no. 1, pp. 130–134, 2000.
- [35] GUPTA, A. and DAVIS, L. S., “Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers,” in *ECCV*, pp. 16–29, Springer, 2008.
- [36] HAMID, R., KUMAR, R. K., GRUNDMANN, M., KIM, K., ESSA, I., and HODGINS, J., “Player localization using multiple static cameras for sports visualization,” in *CVPR*, 2010.
- [37] HAMID, R., MADDI, S., JOHNSON, A., BOBICK, A., ESSA, I., and ISBELL, C., “A novel sequence representation for unsupervised analysis of human activities,” *Artificial Intell.*, vol. 173, pp. 1221–44, 2009.
- [38] HANJALIC, A., “Generic approach to highlights extraction from a sport video,” in *International Conference on Image Processing (ICIP)*, vol. 1, IEEE, 2003.
- [39] HANJALIC, A., “Multimodal approach to measuring excitement in video,” in *International Conference on Multimedia and Expo (ICME)*, vol. 2, IEEE, 2003.
- [40] HANJALIC, A. and XU, L.-Q., “Affective video content representation and modeling,” *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

- [41] HANNON, J., MCCARTHY, K., LYNCH, J., and SMYTH, B., “Personalized and automatic social summarization of events in video,” in *Proceedings of the 16th International Conference on Intelligent User Interfaces*, pp. 335–338, ACM, 2011.
- [42] HARALICK, R. M., “Decision making in context,” *Trans. PAMI*, pp. 417–428, 1983.
- [43] HAREL, J., KOCH, C., PERONA, P., and OTHERS, “Graph-based visual saliency,” *NIPS*, vol. 19, p. 545, 2007.
- [44] HAYS, J. and EFROS, A. A., “Im2gps: estimating geographic information from a single image,” in *CVPR*, pp. 1–8, 2008.
- [45] HE, X., ZEMEL, R. S., and CARREIRA-PERPINDN, M., “Multiscale conditional random fields for image labeling,” in *CVPR*, IEEE, 2004.
- [46] HIGHTOWER, J. and BORRIELLO, G., “A survey and taxonomy of location systems for ubiquitous computing,” *IEEE computer*, vol. 34, no. 8, pp. 57–66, 2001.
- [47] HOASHI, H., JOUTOU, T., and YANAI, K., “Image recognition of 85 food categories by feature fusion,” in *Multimedia (ISM), 2010 IEEE International Symposium on*, pp. 296–301, 2010.
- [48] HOFMANN, T., “Probabilistic latent semantic indexing,” in *ACM SIGIR Conf. on IR*, 1999.
- [49] HOIEM, D., EFROS, A. A., and HEBERT, M., “Geometric context from a single image,” in *ICCV*, vol. 1, pp. 654–661, IEEE, 2005.
- [50] HOIEM, D., EFROS, A. A., and HEBERT, M., “Putting objects in perspective,” *International Journal of Computer Vision*, vol. 80, no. 1, pp. 3–15, 2008.
- [51] HOU, X., HAREL, J., and KOCH, C., “Image signature: Highlighting sparse salient regions,” *Trans. PAMI*, vol. 34, no. 1, pp. 194–201, 2012.
- [52] ITTI, L., KOCH, C., NIEBUR, E., and OTHERS, “A model of saliency-based visual attention for rapid scene analysis,” *Trans. PAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.
- [53] ITU-R, R., “ITU-r bs. 1770-2, algorithms to measure audio programme loudness and true-peak audio level,” *International Telecommunications Union, Geneva*, 2011.
- [54] IVANOV, Y. A. and BOBICK, A. F., “Recognition of visual activities and interactions by stochastic parsing,” *PAMI*, vol. 22, pp. 852–872, August 2000.

- [55] KALOGERAKIS, E., VESSELOVA, O., HAYS, J., EFROS, A. A., and HERTZMANN, A., “Image sequence geolocation with human travel priors,” in *ICCV*, pp. 253–260, 2009.
- [56] KITAMURA, K., DE SILVA, C., YAMASAKI, T., and AIZAWA, K., “Image processing based approach to food balance analysis for personal food logging,” *IEEE International Conference on Multimedia. Proceedings*, pp. 625–630, July 2010.
- [57] KM, F., BI, G., DF, W., and MH, G., “Excess deaths associated with underweight, overweight, and obesity,” *JAMA*, vol. 293, no. 15, pp. 1861–1867, 2005.
- [58] KUMAR, N., BERG, A., BELHUMEUR, P., and NAYAR, S., “Attribute and simile classifiers for face verification,” in *ICCV*, 2009.
- [59] KUMAR, S. and HEBERT, M., “Discriminative random fields: A discriminative framework for contextual interaction in classification,” in *ICCV*, pp. 1150–1157, IEEE, 2003.
- [60] KUMAR, S. and HEBERT, M., “A hierarchical field framework for unified context-based classification,” in *ICCV*, vol. 2, pp. 1284–1291, IEEE, 2005.
- [61] LADD, A. M., BEKRIS, K. E., RUDYS, A., KAVRAKI, L. E., and WALLACH, D. S., “Robotics-based location sensing using wireless ethernet,” *Wireless Networks*, vol. 11, no. 1-2, pp. 189–204, 2005.
- [62] LAPTEV, I., MARSZALEK, M., SCHMID, C., and ROZENFELD, B., “Learning realistic human actions from movies,” in *CVPR*, 2008.
- [63] LE, Q., ZOU, W., YEUNG, S., and NG, A., “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *CVPR*, 2011.
- [64] LE, Q. V., ZOU, W. Y., YEUNG, S. Y., and NG, A. Y., “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *CVPR*, pp. 3361–3368, IEEE, 2011.
- [65] LEE, Y. J., GHOSH, J., and GRAUMAN, K., “Discovering important people and objects for egocentric video summarization,” in *CVPR*, pp. 3–2, 2012.
- [66] LEONARD, J. J. and DURRANT-WHYTE, H. F., “Mobile robot localization by tracking geometric beacons,” *Trans. Robotics and Automation*, vol. 7, no. 3, pp. 376–382, 1991.
- [67] LEPETIT, V. and FUA, P., “Keypoint recognition using randomized trees,” *PAMI*, vol. 28, no. 9, pp. 1465–1479, 2006.

- [68] LI, L.-J. and FEI-FEI, L., “What, where and who? classifying events by scene and object recognition,” in *ICCV*, pp. 1–8, IEEE, 2007.
- [69] LI, L.-J., SOCHER, R., and FEI-FEI, L., “Towards total scene understanding: Classification, annotation and segmentation in an automatic framework,” in *CVPR*, pp. 2036–2043, IEEE, 2009.
- [70] LI, Y., FATHI, A., and REHG, J. M., “Learning to predict gaze in egocentric video,” in *ICCV*, 2013.
- [71] LIU, C., HUANG, Q., JIANG, S., XING, L., YE, Q., and GAO, W., “A framework for flexible summarization of racquet sports video using multiple modalities,” *Computer Vision and Image Understanding*, vol. 113, no. 3, pp. 415–424, 2009.
- [72] LIU, J., KUIPERS, B., and SAVARESE, S., “Recognizing human actions by attributes,” in *CVPR*, 2011.
- [73] LIU, J., LUO, J., and SHAH, M., “Recognizing realistic actions from videos “in the wild”,” in *CVPR*, 2009.
- [74] LOWE, D., “Distinctive image features from scale-invariant keypoints,” *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.
- [75] MANNING, C., RAGHAVAN, P., and SCHÜTZE, H., *Introduction to Information Retrieval*. Cambridge Univ. Press, 2008.
- [76] MARTIN, J., REGEHR, G., REZNICK, R., MACRAE, H., MURNAGHAN, J., HUTCHISON, C., and BROWN, M., “Objective structured assessment of technical skill (osats) for surgical residents,” *British Journal of Surgery*, vol. 84, no. 2, pp. 273–278, 1997.
- [77] MATAS, J., CHUM, O., URBAN, M., and PAJDLA, T., “Robust wide-baseline stereo from maximally stable extremal regions,” *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.
- [78] MATTHEWS, B. W., “Comparison of the predicted and observed secondary structure of t4 phage lysozyme,” *Biochimica et Biophysica Acta (BBA)-Protein Structure*, vol. 405, no. 2, pp. 442–451, 1975.
- [79] MENDI, E., CLEMENTE, H. B., and BAYRAK, C., “Sports video summarization based on motion analysis,” *Computers & Electrical Engineering*, vol. 39, no. 3, pp. 790–796, 2013.
- [80] MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T., and GOOL, L., “A comparison of affine region detectors,” *IJCV*, vol. 65, no. 1, pp. 43–72, 2005.

- [81] MIKOLAJCZYK, K., TUYTELAARS, T., SCHMID, C., ZISSERMAN, A., MATAS, J., SCHAFFALITZKY, F., KADIR, T., and VAN GOOL, L., “A comparison of affine region detectors,” *International journal of computer vision*, vol. 65, no. 1-2, pp. 43–72, 2005.
- [82] MOORE, D. and ESSA, I., “Recognizing multitasked activities from video using stochastic context-free grammar,” in *AAAI*, 2002.
- [83] MURPHY, K., TORRALBA, A., FREEMAN, W., and OTHERS, “Using the forest to see the trees: a graphical model relating features, objects and scenes,” *NIPS*, vol. 16, pp. 1499–1506, 2003.
- [84] NAVON, D., “Forest before trees: The precedence of global features in visual perception,” *Cognitive psychology*, vol. 9, no. 3, pp. 353–383, 1977.
- [85] NICHOLS, J., MAHMUD, J., and DREWS, C., “Summarizing sporting events using twitter,” in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, pp. 189–198, ACM, 2012.
- [86] NIEBLES, J., WANG, H., and FEI-FEI, L., “Unsupervised learning of human action categories using spatial-temporal words,” *IJCV*, 2008.
- [87] NINASSI, A., LE MEUR, O., LE CALLET, P., and BARBBA, D., “Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric,” in *ICIP*, vol. 2, pp. II–169, 2007.
- [88] NORONHA, J., HYSEN, E., ZHANG, H., and GAJOS, K. Z., “Platemate: crowdsourcing nutritional analysis from food photographs,” *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pp. 1–12, 2011.
- [89] OH, S., HOGGS, A., TUREK, M., and COLLINS, R., “Content-based retrieval of functional objects in video using scene context,” in *ECCV*, 2010.
- [90] OLIVA, A. and TORRALBA, A., “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [91] OTSASON, V., VARSHAVSKY, A., LAMARCA, A., and DE LARA, E., “Accurate GSM indoor localization,” in *UbiComp*, pp. 141–158, 2005.
- [92] POWERS, D. M., “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” 2011.
- [93] PRIYANTHA, N. B., CHAKRABORTY, A., and BALAKRISHNAN, H., “The cricket location-support system,” in *Int. Conf. Mobile computing and networking*, pp. 32–43, ACM, 2000.

- [94] RABINOVICH, A., VEDALDI, A., GALLEGUILLOS, C., WIEWIORA, E., and BELONGIE, S., “Objects in context,” in *ICCV*, pp. 1–8, IEEE, 2007.
- [95] ROBERTSON, D. P. and CIPOLLA, R., “An image-based system for urban navigation.,” in *BMVC*, pp. 1–10, 2004.
- [96] SALTON, G., *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, 1971.
- [97] SAZONOV, E., SCHUCKERS, S., LOPEZ-MEYER, P., MAKEYEV, O., SAZONOVA, N., MELANSON, E. L., and NEUMAN, M., “Non-invasive monitoring of chewing and swallowing for objective quantification of ingestive behavior,” *Physiological Measurement*, vol. 29, pp. 525–541, Apr. 2008.
- [98] SCHINDLER, G., BROWN, M., and SZELISKI, R., “City-scale location recognition,” in *CVPR*, pp. 1–7, 2007.
- [99] SCHROTH, G., HUITL, R., CHEN, D., ABU-ALQUMSAN, M., AL-NUAIMI, A., and STEINBACH, E., “Mobile visual location recognition,” *Signal Processing Magazine, IEEE*, vol. 28, no. 4, pp. 77–89, 2011.
- [100] SENATOR, T. E., BETTADAPURA, V., and OTHERS, “Detecting insider threats in a real corporate database of computer usage activity,” in *Proceedings of 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, August 2013.
- [101] SHAPOVALOVA, N., RAPTIS, M., SIGAL, L., and MORI, G., “Action is in the eye of the beholder: Eye-gaze driven model for spatio-temporal action localization,” in *NIPS*, pp. 2409–2417, 2013.
- [102] SHI, J. and TOMASI, C., “Good features to track,” in *IEEE Computer Vision and Pattern Recognition (CVPR)*, pp. 593–600, IEEE, 1994.
- [103] SINGHAL, A., LUO, J., and ZHU, W., “Probabilistic spatial context models for scene content understanding,” in *CVPR*, vol. 1, pp. 1–235, IEEE, 2003.
- [104] SMITH, R., “An overview of the tesseract ocr engine,” in *icdar*, pp. 629–633, IEEE, 2007.
- [105] SONNENBURG, S., RÄTSCH, G., SCHÄFER, C., and SCHÖLKOPF, B., “Large scale multiple kernel learning,” *The Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.
- [106] STRAT, T. M. and FISCHLER, M. A., “Context-based vision: recognizing objects using information from both 2d and 3d imagery,” *Trans. PAMI*, vol. 13, no. 10, pp. 1050–1065, 1991.
- [107] SUDDERTH, E. B., TORRALBA, A., FREEMAN, W. T., and WILLSKY, A. S., “Learning hierarchical models of scenes, objects, and parts,” in *ICCV*, vol. 2, pp. 1331–1338, IEEE, 2005.

- [108] TANG, A. and BORING, S., “# epicplay: crowd-sourcing sports video highlights,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1569–1572, ACM, 2012.
- [109] THOMAZ, E., BETTADAPURA, V., REYES, G., SANDESH, M., SCHINDLER, G., PLÖTZ, T., ABOWD, G. D., and ESSA, I., “Recognizing water-based activities in the home through infrastructure-mediated sensing,” in *Proceedings of ACM Conference on Ubiquitous Computing (UbiComp)*, pp. 85–94, ACM, 2012.
- [110] THOMAZ, E., PARNAMI, A., BIDWELL, J., ESSA, I. A., and ABOWD, G. D., “Technological approaches for addressing privacy concerns when recognizing eating behaviors with wearable cameras,” *UbiComp*, pp. 739–748, 2013.
- [111] THOMAZ, E., PARNAMI, A., ESSA, I. A., and ABOWD, G. D., “Feasibility of identifying eating moments from first-person images leveraging human computation,” *SenseCam*, pp. 26–33, 2013.
- [112] TJONDRONEGORO, D., TAO, X., SASONGKO, J., and LAU, C. H., “Multi-modal summarization of key events and top players in sports tournament videos,” in *IEEE WACV*, pp. 471–478, IEEE, 2011.
- [113] TONG, X., LIU, Q., ZHANG, Y., and LU, H., “Highlight ranking for sports video browsing,” in *Proceedings of the 13th annual ACM international conference on Multimedia*, pp. 519–522, ACM, 2005.
- [114] TORRALBA, A., “Contextual influences on saliency,” *Neurobiology of Attention*, 2005.
- [115] TORRALBA, A., MURPHY, K. P., and FREEMAN, W. T., “Contextual models for object detection using boosted random fields,” in *NIPS*, pp. 1401–1408, 2004.
- [116] TREISMAN, A. M. and GELADE, G., “A feature-integration theory of attention,” *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.
- [117] TU, Z., CHEN, X., YUILLE, A. L., and ZHU, S.-C., “Image parsing: Unifying segmentation, detection, and recognition,” *International Journal of Computer Vision*, vol. 63, no. 2, pp. 113–140, 2005.
- [118] TURAGA, P., CHELLAPPA, R., SUBRAHMANIAN, V. S., and UDREA, O., “Machine recognition of human activities: A survey,” *IEEE Trans. Circuits Systems for Video Tech.*, October 2008.
- [119] USDA, “Food consumption and demand,” Nov. 2013.
- [120] VAN DE SANDE, K., GEVERS, T., and SNOEK, C., “Evaluating color descriptors for object and scene recognition,” *PAMI*, vol. 32, no. 9, pp. 1582–1596, 2010.

- [121] VISHWANATHAN, S., SUN, Z., THEERA-AMPORNPUNT, N., and VARMA, M., “Multiple kernel learning and the smo algorithm,” *NIPS*, 2010.
- [122] WALLACH, H., “Topic modeling: beyond bag-of-words,” in *ICML*, pp. 977–984, ACM, 2006.
- [123] WANG, H., KLASERR, A., SCHMID, C., and LIU, C., “Action recognition by dense trajectories,” in *CVPR*, 2011.
- [124] WANG, H., ULLAH, M. M., KLÄSER, A., LAPTEV, I., and SCHMID, C., “Evaluation of local spatio-temporal features for action recognition,” in *BMVC*, 2009.
- [125] WANG, H., KLASER, A., SCHMID, C., and LIU, C.-L., “Action recognition by dense trajectories,” in *CVPR*, pp. 3169–3176, IEEE, 2011.
- [126] WANG, H., ULLAH, M. M., KLASER, A., LAPTEV, I., and SCHMID, C., “Evaluation of local spatio-temporal features for action recognition,” in *BMVC*, 2009.
- [127] WANT, R., HOPPER, A., FALCAO, V., and GIBBONS, J., “The active badge location system,” *ACM Trans. on Information Systems*, vol. 10, no. 1, pp. 91–102, 1992.
- [128] WOLF, L. and BILESCHI, S., “A critical view of context,” *IJCV*, vol. 69, no. 2, pp. 251–261, 2006.
- [129] XIONG, Z., RADHAKRISHNAN, R., and DIVAKARAN, A., “Generation of sports highlights using motion activity in combination with a common audio feature extraction framework,” in *International Conference on Image Processing (ICIP)*, IEEE, 2003.
- [130] XIONG, Z., RADHAKRISHNAN, R., DIVAKARAN, A., and HUANG, T. S., “Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP’03)*, vol. 5, IEEE, 2003.
- [131] XU, C., WANG, J., LU, H., and ZHANG, Y., “A novel framework for semantic annotation and personalized retrieval of sports video,” *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 421–436, 2008.
- [132] YANG, S., CHEN, M., POMERLEAU, D., and SUKTHANKAR, R., “Food recognition using statistics of pairwise local features,” in *CVPR*, 2010.
- [133] YATANI, K. and TRUONG, K. N., “BodyScope: a wearable acoustic sensor for activity recognition,” *UbiComp ’12: Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 341–350, 2012.

- [134] ZAMIR, A. R. and SHAH, M., “Accurate image localization based on Google maps street view,” in *ECCV*, pp. 255–268, 2010.
- [135] ZHANG, J., MARSZALEK, M., LAZEBNIK, S., and SCHMID, C., “Local features and kernels for classification of texture and object categories: A comprehensive study,” *IJCV*, vol. 73, no. 2, pp. 213–238, 2007.
- [136] ZHAO, Z., JIANG, S., HUANG, Q., and ZHU, G., “Highlight summarization in sports video based on replay detection,” in *International Conference on Multimedia and Expo*, pp. 1613–1616, IEEE, 2006.