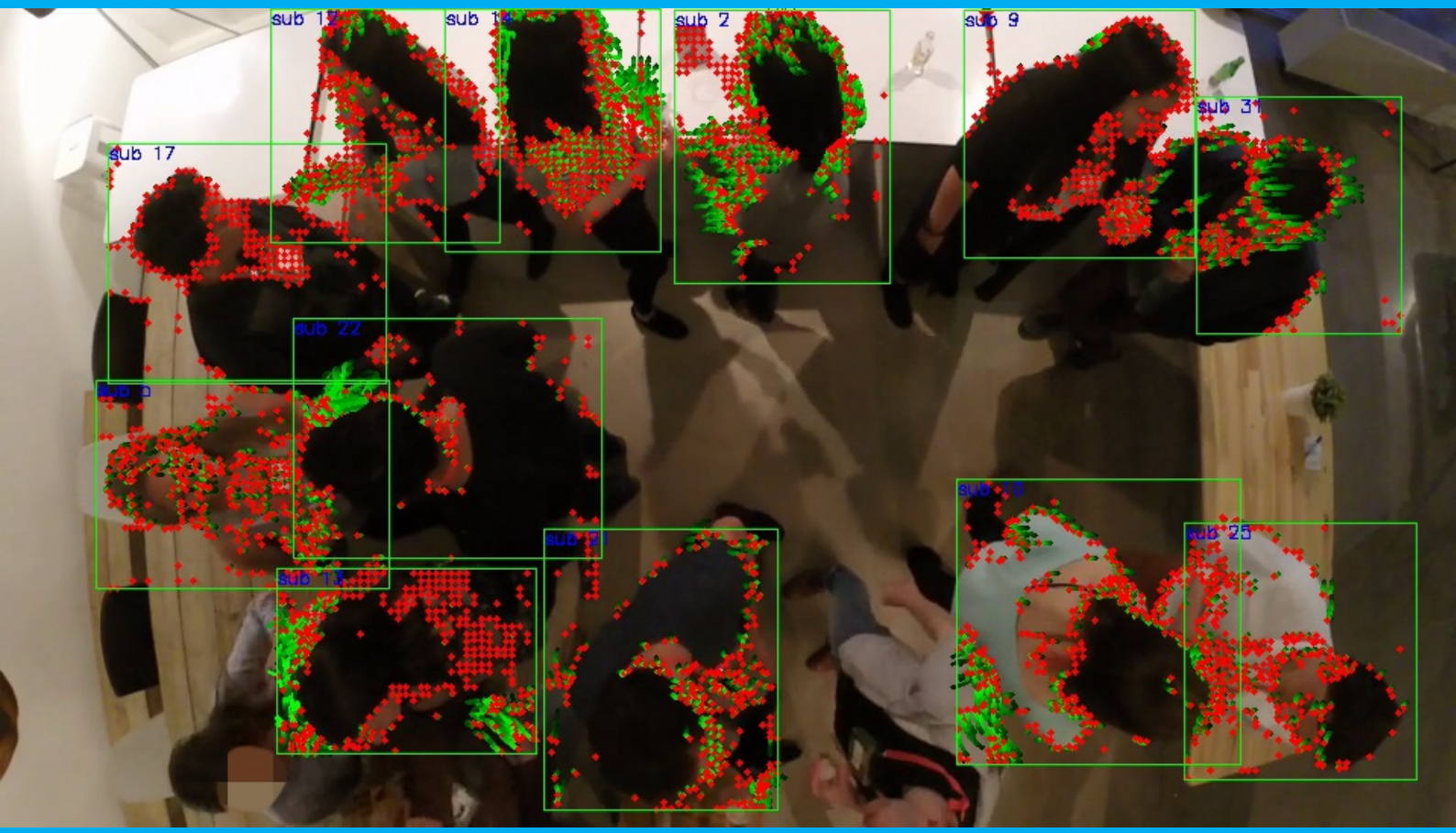
Drinking Behaviour Detection

Using both Static and Dynamic information

X. Teng



TU Delft

Drinking Behaviour Detection

Using both Static and Dynamic information

by

X. Teng

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Tuesday June 18, 2019 at 15:00 AM.

Student number:	4574060	
Project duration:	June 1, 2018 – June 18, 2019	
Thesis committee:	Assoc. Prof. dr. ir. H. Hung,	TU Delft, supervisor
	Assoc. Prof. dr. ir. Frans A. Oliehoek,	TU Delft
	Dr. L. Cabrera-Quiros,	TU Eindhoven, daily supervisor

This thesis is confidential and cannot be made public until June 18, 2019.

Preface

This basis for this research originally stemmed from my passion for developing better methods of drinking behavior detection in a crowded setting with only top-down perspective videos available. As the social signal processing is one of the elements in human-machine interactions analysis, classifying and localizing the action both in spatial and temporal domain draws enormous attention in recent decades. How will we achieve this target? It is my passion to not only find out, but to develop a general approach that can also be used for other future similar researches.

In truth, I could not have achieved my current level of success without a strong support group. First of all, my parents and my girlfriend, who supported me with love and understanding. And secondly, my committee members, each of whom has provided patient advice and guidance throughout the research process. Thank you all for your unwavering support.

X. Teng
Delft, June 2019

Contents

1 Introduction

Human action recognition in videos [5, 7, 25, 44] has gained a lot of attention in the past decade as video becomes a universal source of information and computational power has developed dramatically. Its wide range application in: surveillance systems[29], content-based video retrieval, health-care monitoring system, human-computer interaction, gaming, and socially perceptive interpretation make it be one of the most active research fields. Social Signal Processing, a domain focused on understanding of social interactions through machine analysis, is the interested field of this research. The aim of this work is to develop a method for drinking action detection and temporal localization in videos recorded in a top-down perspective.

Drinking action plays a crucial role in social activities. Researches [3, 6, 10, 42] about the relations between drinking action and other social aspects such as attractiveness, social approval and facilitation have been done in the past years. Empirical evidence[14, 16, 21, 46] suggests non-conscious mimicry creates affiliation, and affiliation can be expressed through non-conscious mimicry. Undeniable, the mimicry in drinking behaviors such as drinking pace, drinking duration, and sip rate is one of the key elements in this analysis[16]. However, no matter which type of analysis we do, they all based on atomic drinking moment detection.

In this work, machine learning algorithms are used to extract drinking action features in crowded mingle scenarios. As motion and shape information are two key elements for human action localization, we used two types of feature descriptors to extract their information independently. After that, a hybrid model that can take both insights from motion and appearance cues into consideration in the final decision-making process is implemented. In the rest of this work, we used `dynamic` to represent the motion-based features and `static` to describe the shape-based features. According to our literature review, there are not many works doing action recognition based on both motion and shape features. Laptev and Patrick [25] is an early attempt with using this methodology. The keyframe classifier combined with the space-time classifier is used for human action localization. In recent years, Chéron et al. [9] treated motion and shape as two independent streams as the inputs for two neural networks and fusing their results to detect human action.

In the rest of this chapter, the motivation of this research is presented, After that, the research questions together with the main hypotheses are summarized and an elaboration on the contributions is given. Subsequently, an outline of the rest of the thesis is given.

1.1. Motivation

In this work, we focus on the drinking action detection which is the foundation of all mentioned researches. We propose a specific classification model that capable to determine a precise starting and ending point of a drinking action. We assume the interactions of drinking behavior among participants can also be used as an indication of attractiveness[3, 6]. Attractiveness level determination among participants is one of the most important tasks in the social interactions analysis of mingle related activities.

In fact, the exclusion or engagement level of a person in a group is the key element to show the attractiveness of a person to the others. However, they are hard to measure. Therefore, we start with the detection of mimicking in drinking behaviors. As mimicking in drinking behaviors is detectable, the features of attractiveness could be quantified by using the correlation relationship between them and mimicry.

1.2. Dataset

The dataset used in this project for model evaluation is called MatchNMingle[4]. It is a multi-sensor resource normally used for the analysis of social interactions and group dynamics in-the-wild. Only the video source recorded by cameras mounted on the ceiling is used. In total, there are 92 participants were recorded for 30 minutes at 20 FPS.

Figure.1.1 gives a more intuitive perception. The `Region Of Interested` (ROI) noted as the green bounding boxes is given by the dataset. So action localization in the space domain is not a task of this report. The red dots and green curves are explained in other chapters.

This dataset has several characteristics. Firstly, participants stand close to each other. It results in strong occlusions among ROIs. The cross-contamination between subjects is another problem caused by this. Moreover, the drinking detection in this dataset becomes more challenging due to the individual variations of participants in the orientation of participants and their motion styles.

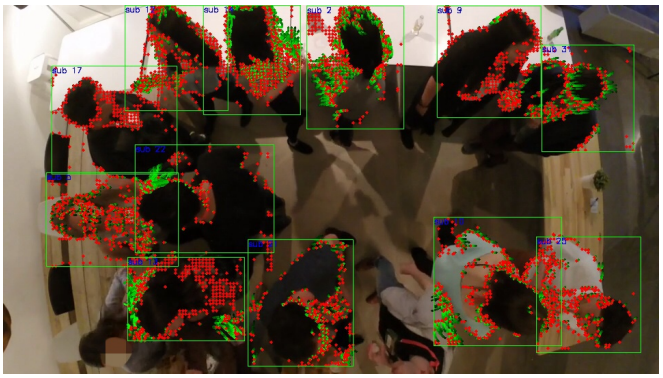


Figure 1.1: a top down perspective of the video source offered by the MatchNMingle dataset

1.3. Problem Specification

The goal of this project is to design a classification model that is capable to determine the precise starting and ending point of each drinking action happens in the recorded videos given by MatchNMingle. So, what are the problems that we need to solve to achieve this goal? We assume that both motion and shape can be used as the inputs of the model simultaneously for drinking moments detection. The first task is to find out how to synchronize both parts that belong to a different domain. Meanwhile, we have to figure out what are the features and descriptors that we can use for motion and shape information extraction with the largest descriptive power acquired. it relies on the extraction techniques that we choose to perform this task. The algorithms which are suitable for the features of the dataset mentioned in Section 1.2 are considered.

Although the computational power has developed dramatically in the past decade, It is still computationally expensive if video-based analysis involved. To reduce the requirement of computation, a proper sampling strategy is needed. This is essential while a frame-level classification is implemented.

1.4. Research Questions

The central research topics of this research are listed as questions as follows,

1. *Main*: Will the static shape in frame-level provide extra information besides motion cues while motion-level is low in drinking action detection in a crowded environment?

2. *Sub:* How to fuse these two independent models together?

3. *Sub:* How to synchronize both parts, in other words, how to align both the static and dynamic information in a fixed length window. And if the synchronized strategy respect to a fixed length window is suitable in action localization in the temporal domain?

4. *Sub:* What is the sampling strategy of each part? How to measure the effectiveness of the sampling strategy. In other words, how to check if only the samples with a higher descriptive power are selected by the strategy.

1.5. Hypothesis Formulation

When using a scientific method to solve problem or question, one formulates educated guess or hypothesis which is a possible answer to the problem or question. The hypothesizes that we drawn in this section help

us to find out if the proposed model is working in this project. Also, it helps to answer questions mentioned in Section 1.4.

Both dynamic and static information are important for drinking action detection
In general, human action information can either be encoded by its motion or appearance. Experiments in Laptev and Patrick [25] has claimed that the keyframe(shape) classifier appears to be complementary to the space-time(motion) classifier. Intuitively, the motion-based features will have less descriptive power while the motion level is low. This happens while the hand region of participants is actually contacting with their face or lips. However, the appearance-based features have the largest descriptive power in this stage as the shape looks most like drinking.

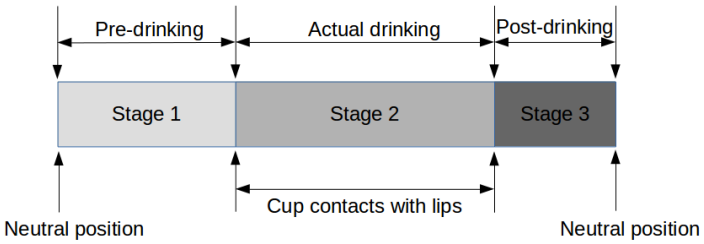


Figure 1.2: the definition of a complete drinking action

Figure 1.2 shows a typical drinking action consists of three stages. In the beginning, people start to move their hand from a neutral position towards their lips, the dominated motion happens around hand and head followed by the second stage when the cup touches their lips. After that, movements occur around the hand and head again while people start to move their cup down at the ending stage of the whole drinking action. Both motion and shape are important for classification problem respect to a fixed window size as we are not sure which stage is included in the current window.

(a) stage 1 (b) stage 2 (c) stage 3

Figure 1.3: the corresponding figures of the stages mentioned in Fig 1.2

Figure 1.3 given the corresponding pictures of the stages mentioned in Figure 1.2. Both Figure 1.3a and 1.3c are descried as a neutral position while the figure 1.3b is defined as the actual drinking as the cup contacts with the lips of the participant.

High performance will be achieved using a hybrid model
Why a hybrid model might give higher performance than each of them? Firstly, we have to be clear about how drinking action is defined in this project. It is clearly interpreted in Figure 1.2. The definition of it in MatchNMingle[4] is the duration that subject's hand starts to move from the neutral position until it ends at also the neutral position. The `actual drinking` starts when the cup is touching the subject's lips. This is also the stage where the appearance information turns dominant, as only a small fraction of motion is happening while people are drinking (Hypothesis 1.5.3). The motion part is dominant at the beginning and ending stage of the drinking action. By observation, the duration of the actual drinking has almost the same length with the combination of the beginning and ending stage or even longer. Therefore, both parts are equally important in classification respect to a window. Furthermore, It is not clear which stage is included in the cutout section after sliding window operation. For example, the motion part will be more important if only the start or end section of the action is included in the current window.

High performance fusion strategy
A fusion strategy based on the global motion level and the effective motion in a window is effective. In fact, fusion is the process of finding the contribution weights of all features in the final decision model. Here, I

propose that the contribution weights are relating to two ratios. the first ratio is described as the motion level over the entire period and the second one is the percentage of the motion that is really used for decision making in the current window, For instance, in drinking behavior detection, only the motion related to hand and shoulder will count for decision making, the motion of the rest part of the body is considered as noise.

Motion level decreases as people start to drink
We expect that the calculated motion level will decrease as people start to drink. Intuitively, people trends to be still while they begin to drink to avoid spilling of beverage. More details shown in Figure. 1.2.

1.6. Contributions

There are not many works conducted on drinking action detection in a real-world setting. Previous works have either focused on a general human action classification and localization or drinking action detection in a laboratory setting. Besides, appearance information draws quite a few attentions on human action recognition in a real-world setting environment because of its large variation over different objects and actions. However, we strongly believed that drinking is an action that could be recognized by its natural description. It means the shape of drinking itself might also largely contribute to its detection. On the other side, motion is always indispensable to human action detection. Hence, we claim that both motion and shape are equally important for drinking action detection in this project.

Main contributions of this work are:

- Perform the task of drinking action detection in a crowded environment in a top view.
- Create a fusion model which can take the advantages of the strength from both shape and motion information of drinking action.
- Use motion intensity for a fusion model design for drinking action detection.

1.7. Outline

The rest of the paper is organized as follows. Chapter 2 elaborates the related work in details while Chapter 3 contains the proposed technique. The experimental setup and their results are discussed in Chapter 4.5 and 5.3 respectfully followed by conclusion.

2

Related Work

Normally, human action detection and localization both in spatial and temporal space are studied together[9, 20]. Here we focus on temporal localization only with using the given ROI(bounding boxes) in MatchNMingle. The objective of temporal action localization is both to classify and identify temporal extents of actions in videos. Spatio-temporal features are successful representation for action recognition in videos. There is a large amount of methods[27, 36, 39] for extracting local spatio-temporal features in videos. Kläser et al. [20], Laptev and Lindeberg [22] have introduced spatio-temporal interest points, which are an extension of the Harris detector from image to video. some other extensions of image descriptors, such as 3D-SIFT([40]), HOG3D([19]), extended SURF([49]), and Local Trinary Patterns([51]). Nevertheless, some other people may have a different opinion. Dollár et al. [13] argued that direct 3D counterparts to 2D interest point detectors are inadequate for the detection of spatio-temporal feature points since true spatio-temporal corners are quite rare. Also, 2D space domain and 1D time domain in videos show different characteristics. It is, therefore, more intuitive to handle them in a different manner than to detect interest points in a joint 3D space. We have agreed on this opinion and treated them independently in our work.

A survey done by Borges et al. [2] separates Video-based Action Detection methods into three main groups, namely, appearance-based, motion-based, and hybrid methods. Appearance-based techniques are a specific case of action detection in still images and are used for example edge information. Motion-based methods consider temporal information for the definition of features defining an action, in particular, the movements of the hands. Hybrid methods use combinations of the previous two classes. Our literature review is conducted also based on this division.

2.1. Appearance-based

Delaitre et al. [12] indicates that a natural description of many still images is provided by the action itself. Shape representations in terms of histograms of image gradients have shown excellent performance on object recognition problem [11, 23, 30]. However, the global histograms[38] do not suit for our case due to a large variation in both appearance and orientation of participants, a better approach supported in consists of computing histograms over local image regions. Now, the problem turns to be where those image regions should be and what are the size of them. Laptev [23] using the given training dataset for solving these problems. They choose the position and the shape of histogram features to minimize the training error. And then AdaBoost[15] is used to select histogram features and to learn an object classifier. This method performs well while a set of scale and position normalized object images with similar views. It is not true in our case. Another proposal for appearance information extraction is through BoF. Previous works[13, 24, 39, 48] trend to use BoF together with space-time features for action detection. Here, it is only used for extracting the shape information. No matter which uses of BoF, it has its limitation. Ullah et al. [45] indicate that it needs to balance a trade-off between discriminative power and the invariance needed to overcome irrelevant variations in the video. The solution offered by this paper is by integrating non-local cues available at the region-level of a video to improve discriminative power. However, this might not be a solution for us. Our research topic is specified to be a temporal action localization problem as ROI is given in MatchNMingle [4]. Moreover, the scene presented in the video is crowded. Overlapping over different ROI is inevitable.

2.2. Motion-based

The motion cues here in our work are used to overcome the limitation of using region-level based features. Researches[33, 47] show good results of leveraging the motion information of trajectories. There are several ways could be used for generating trajectories. Messing et al. [35] extract feature trajectories by tracking Harris 3D interest points with a KLT tracker [31] while [43] compute trajectories by matching SIFT descriptors between two consecutive frames. we are using the method mentioned in [47] as he has already done a comparison among different approaches. A new feature called Trajectory-Set (TS) is proposed in [34] compared with the DT defined in [47]. It encodes only trajectories around densely sampled interest points, without any appearance features as they think that motion information is discriminative enough for classifying different actions. Besides, the way of aggregating trajectories is different. In TS, the frame is divided into cells of M×M pixels and group the trajectories in each cell.

Back to our topic with the specific MatchNMingle dataset, Cabrera-Quiros et al. [5] is already one of the implementations of gesture detection on the MatchNMingle dataset. It shows a method used bags of dense trajectories combined with Multiple Instance Learning via Embedded Instance Selection (MILES) to achieve gesture detection in a real-world setting. To compare with gesture detection, the appearance-based analysis seems to be more important as the natural description of the static drinking frames are more meaningful than gesture contained frames. In our approaches, we suggest to extract appearance information individually and integrating it with the motion part instead of using space-time features directly. Experiments in [25] claims that the keyframe classifier appears to be complementary to the space-time classifier. In our research, we suggest that the still image classifier will complementary to the motion-based classifier for a drinking action in a fixed length window.

Another research [20] that is specifically focused on human action localization in videos has also worked on datasets with a crowded, dynamic environment, partial occlusion and cluttered background, such as the *Coffee&Cigarettes* dataset and the new *Hollywood-Localization* dataset. HOG3D features are used for human action localization in spatial space. The sliding window approach extracts descriptors at varying locations and scales is an outstanding point in their method. When training the sliding window classifier, the temporal slices are aligned with the ground-truth begin and end timestamps of the action. At test time, a sliding window with multiple temporal scales is used to localize actions.

2.3. Hybrid

An early attempt on fusion is mentioned in the work of Laptev and Perez [26]. It found the most "active" parts of the space-time classifier are associated with the regions of hand motion while the learned keyframe classifier shows most of its selected (spatial) features being located at the head and the face regions on the keyframe. They appear to be complementary with each other. Their fusion strategy is based on this property. On the other hand, a combination of motion and appearance information learned by two separate CNN networks has been proposed in [37]. Moreover, Chéron et al. [9] is another recent research fusing both appearance and optical flow by using a `recurrent localization network` (RecLNet). It has successfully modeled the temporal structure of actions on the level of person tracks and is based on two-layer `gated recurrent units` (GRU) applied separately to two streams, i.e. appearance and optical flow streams. After that, a last fully-connected layer with softmax converts the recurrent output to an action probability for all actions. It shows to treat appearance and motion independently working for temporal localization of human action. Recently, Ullah and Jaffar [44] proposed a collaborative approach based on holistic and motion information for human activity localization and recognition.

2.4. Others

Besides, some other types of approaches in action localization are also studied. Jain et al. [18] encode the presence of object categories for action classification and localization. It shows that they are semantically relevant, especially when the actions interact with objects. For instance, the presence of cups must be relevant with drinking action detection, however, it is not enough for drinking detection if only the information of the presence of cups is given. The selection from multiple objects appeared in the scene is necessary. Action recognition is a popular and well-studied topic. However, more recently CNN and RNN-based methods [1, 32, 41, 52] have shown gains for action localization.

3 Methods

3.1. Method motivation

Jain et al. [18] revealed that object-action relations are generic, which allows transferring these relationships from the one domain to the other. Objects, when combined with motion, improve the state-of-the-art for both action classification and localization. It is undeniable that both shape and motion are crucial in action recognition. Previous works try to use spatial-temporal features to encode both information simultaneously and train a classifier based on these 3D features. we do not fully agree with this type of methods as we believe spatial and temporal features have different characteristics. It is more valuable if we can treat them independently. Chapter 2.4 has already given a large group of alternative approaches for motion information extraction. Experimental results have shown DT is one of the optimal solutions for trajectory information retrieving. Delaitre et al. [12] has shown the value of appearance in human action recognition, especially for temporal action localization. `Boosted Histograms`[23] and `BoF` are normally used for appearance information extraction when images contain a large variation. In this project, we select BoF for shape information extraction due to the orientation of subjects are also variations in a large scale. None of them can achieve a pursued performance on our dataset with people stand fairly close with each other and cameras recorded in a top-down perspective. Both of them have their corresponding limitation in action detection in this situation. To fuse them with a strategy that can take advantage of both strengths is needed. Moreover, Laptev et al. [27] has proved that both parts appear to complement with each other especially for drinking action detection. Therefore, A hybrid solution for drinking action detection is one of the optimal choices. Hauptmann [17] is another example about fusion.

3.1.1. Why BoF instead of HOG ?

Several difficulties appear if HOG extraction is implemented in this case. First of all, the size of the bounding boxes is not consistent. Therefore, the number of descriptors extracted from the bounding boxes are also various. Secondly, the orientation of subjects in bounding boxes is different. To answer the first question, we have to resize all bounding boxes to the mean size of all images before feature extraction. It seems that the problem is solved, but we think it is necessary to have a certain aspect ratio regarding to all images. An extra doubt is that if HOG feature is scale invariant or not. Another suggestion is to use a method that combines AdaBoost learning with local histogram features mentioned in [23]. This is what they call `boosted histograms`.

We can also set the same number of cells for all images instead of resizing them. For example, dividing images into 5×10 cells, the dimensions of the cells for each image will be different. The related consequence is not clear. In contrast, BoF selects the feature patches in different positions of an image and use them to represent the current image. It reduces negative influence on the experiment's implementation caused by differences in image dimensions. Therefore, BoF is selected after all these analysis and considerations.

3.1.2. Why extra BoF?

Although a drinking action can be detected by utilizing the trajectories extracted from DT, the start and end points of each drinking can not be specified if only motion data is available. Nevertheless, appearance-related

information has already been considered into the method of DT, then why we still need BoF to obtain static information? The reasons are given below:

- As shown in Figure 3.1, only the regions around interesting points are analyzed. The definition of the interesting points here is the start point of all selected trajectories. Trajectories are classified as invalid when their variances are out of the range. It also leads to the removal of the points that belong to them. However, it is not clear if the appearance information around these points is influential in decision making. For instance, if these points are around a cup or hand.

- For subject independent, the shape and orientation of subjects are diverse. This leads to a poor performance by using motion cues only. Another approach that can ignore spatial information is needed.

- In MILES, Classifier selects meaningful instances through the training process. In other words, some of the feature points are dropped. However, the regions around them can be interesting if there exists a method making decisions via learning still images.

3.2. Motion

Motion is defined as "the change in position of an object with respect to its surroundings in a given interval of time". Two types of motion descriptors are used to describe the motion level of a frame or a window. The first one is the distance of trajectories defined in Section 3.2.1 as `l`. The motion level of a window is the mean distance of all trajectories that either start or end in this window. And the motion level of a frame is the sum of the optical flow magnitude happening around the selected interesting feature points. The interested regions are shown in Figure 3.1.

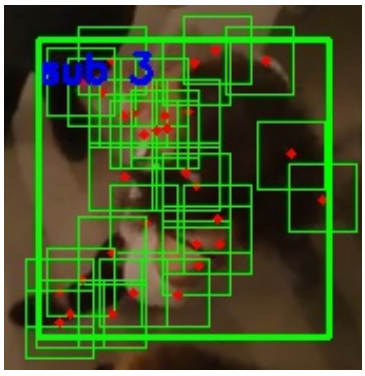


Figure 3.1: Appearance-based feature extraction on a static image

There are two main functions of motion related elements in this project. The most important usage is acting as descriptors of drinking action. So, drinking action could be identified according to the motion trajectories appearing around hand and arm. Another usage of them is for weights calculation. We assume that the frames with less motion involved will contribute more in final decision making. The normalized motion level mentioned in the previous paragraph is regarded as the weight of each frame in a window.

3.2.1. The features of Dense Trajectories

The first 9 elements are information about the trajectory and the other 5 elements are descriptors:

`f:`	The trajectory ends on which frame
`s:`	The trajectory is computed on which scale
`l:`	The length of the trajectory $\left(\sum_{i=1}^{n-1}\sqrt{(x_{i+1}-x_i)^2+(y_{i+1}-y_i)^2}\right)$
m_x:	The mean value of the x coordinates of the trajectory
m_y:	The mean value of the y coordinates of the trajectory
σ_x^2:	The variance value of the x coordinates of the trajectory
σ_y^2:	The variance value of the y coordinates of the trajectory
x_o:	The value of x coordinates of the starting point of trajectory
y_o:	The value of y coordinates of the starting point of trajectory
`Trajectory:`	2x[trajectory length] (default 40 dimension)
`HOG:`	8x[spatial cells]x[spatial cells]x[temporal cells] (default 96 dimension)
`HOF:`	9x[spatial cells]x[spatial cells]x[temporal cells] (default 108 dimension)
`MBHx:`	8x[spatial cells]x[spatial cells]x[temporal cells] (default 96 dimension)
`MBHy:`	8x[spatial cells]x[spatial cells]x[temporal cells] (default 96 dimension)

Therefore, the total dimension of each feature is 9+40+96x3+108 = 455. Some of them are spatial related. Only the spatial unrelated one are chosen as the feature for subject independent classification. More details are discussed in the following sections.

3.2.2. Motion Vs Activity

Gmotion, standing for the Global Motion of trajectories, belong to each frame. In Section 3.4.2, it is defined as the mean of trajectories' length. In this section, we want to find out if there is exactly a correlation relationship between the global motion and the drinking activities. This is also our fundamental hypothesis mentioned in Section 1.5. Figure 3.2 shows a relation between the global motion level and the ground truth when the testing data from the first fold of subject 3 used. The binary result of the blue line represents the ground truth while the purple curve is the measured global motion.

In general, there are two types of motion behaviors could be observed in the graph while drinking activity is taking place. The motion level at a rather low position, such as bar 2,4,8,9 or there is a valley observed during the drinking duration, such as bar 1,6,7,8,9. Besides, bar5 could be viewed as fault detection if we take the result shown in Figure 5.2 as reference. In bar5, the motion level starts to increase while drinking motion starts. The low-level motion could be explained as people trends to move less while starting to drink. And the valley could be interpreted as the motion level is high both for the starts and ends of drinking activity and low when people are really drinking.

However, for some narrow bars, such as 2,4, and 8, the motion various may still seem problematic. It is because only partial of the current action are included in its corresponding window while sliding window is used on the original dataset. It could be seen in Figure 6.1.

Figure 3.3 shows the different drinking and motion status of a subject. The upper part gives the appearance information and the lower part presents the motion information through the corresponding trajectories. For the upper part, Figure 3.3a and 3.3c are the frames annotated as non-drinking in the given ground truth while Figure 3.3b is labeled as drinking. The lower pictures with the selected feature points(red dots) and their corresponding trajectories(green curves) are placed under their original images. The trajectory is defined as the Euclidean movement over the past 20 frames. The average trajectory length is short while people are motionless shown in Figure 3.3d and it is relatively long while the subject is moving shown in Figure 3.3f. 3.3e shows the trajectories' distribution while an actual drinking action is undertaken. Most of the trajectories occur around the hand and arm of the subject.

Previous sections have already discussed a point that motion and shape appear to be complementary to each other. Here is some evidence. It is hard to say whether the subject in Figure 3.3c is drinking or not if

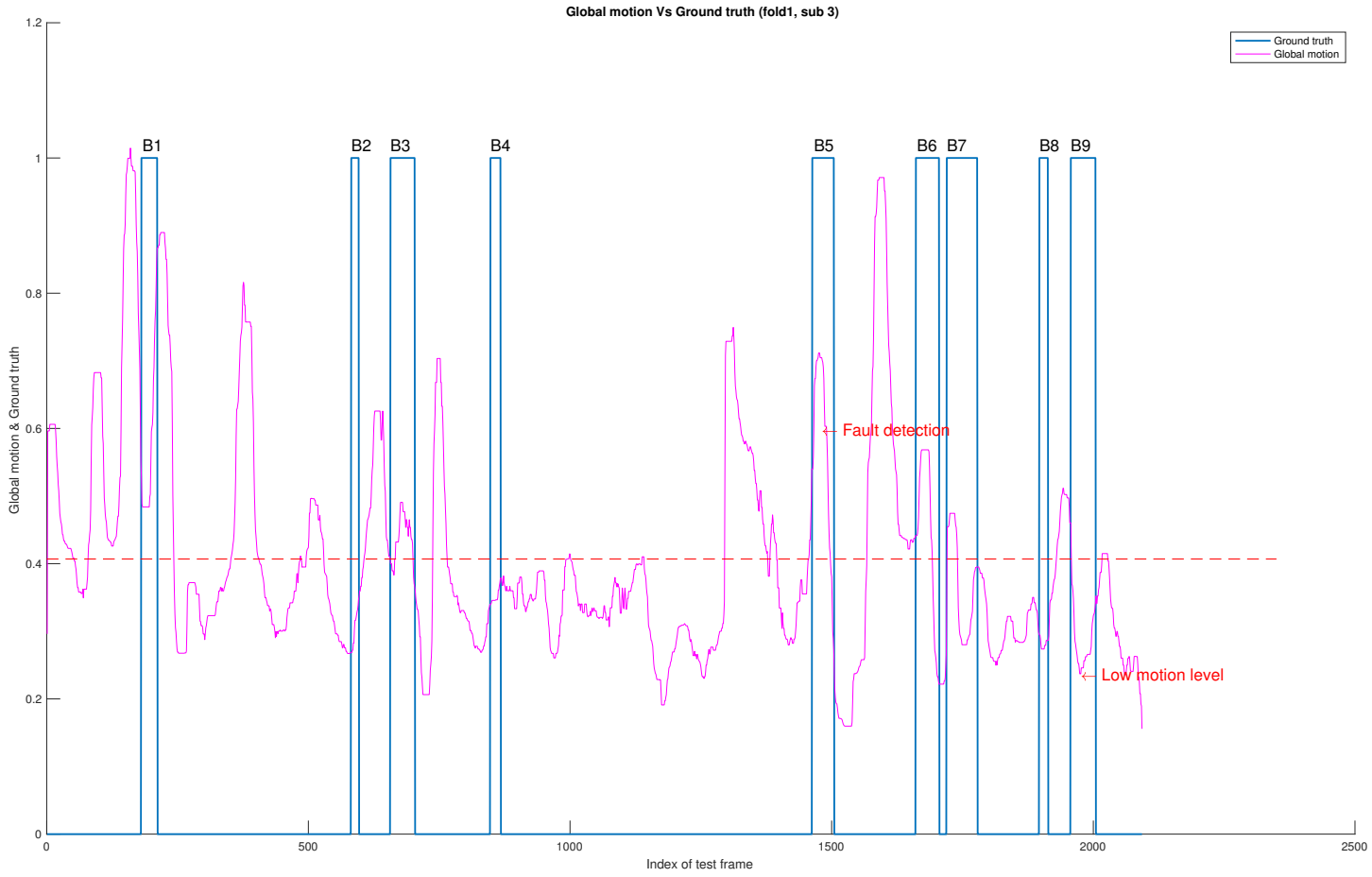


Figure 3.2: relation between ground truth and global motion level

we only take the appearance information into consideration. However, the trajectories shown in Figure 3.3f indicates that the motion level for the current frame is relatively high. In other words, it is most likely that the person is not drinking as we assumed people try to reduce the overall body motion level if they are starting to drink. On the other side, it is also hard to judge if only a few motion information available shown in Figure 3.3d. However, we can see from the original image shown in Fig 3.3a is not drinking.

3.3. Appearance

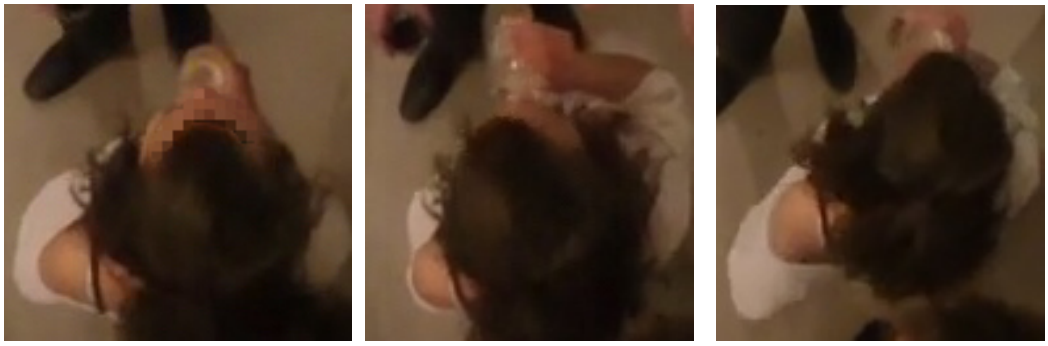
In this segment, the technique used for feature extraction from still images is elaborated. It is wise to learn about the dataset characteristics before exploring the methods used for feature extraction. In MatchNMingle, both the appearance and orientation of subjects vary tremendously. Besides, people's drinking behaviors are also diverse. It is necessary to have a robust method which can be insensitive to spatial information.

3.3.1. Interested regions

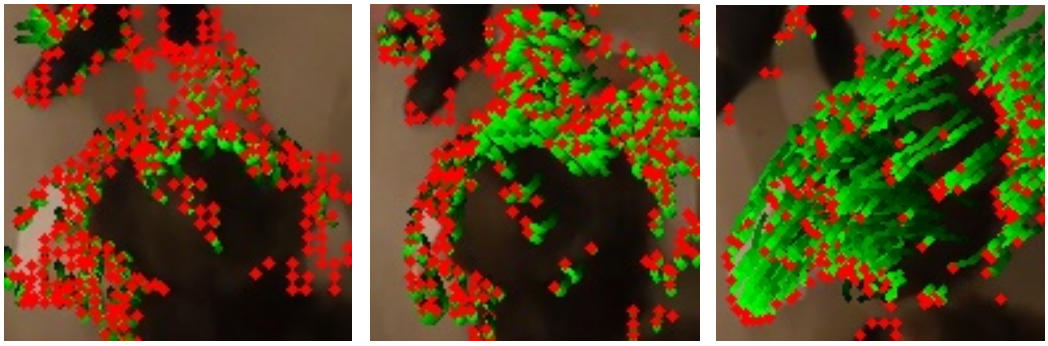
The "interested regions" in this project is defined as the area inside of the annotated bounding boxes. Figure 3.1 shows the feature points and their corresponding blocks, which were collected on the first scale.

3.3.2. Subject orientation

A comprehensive study on the dataset needed before the actual methodology selection process or any experiment starts. We find that the MatchNMingle dataset has an enormous variation both in the appearance and



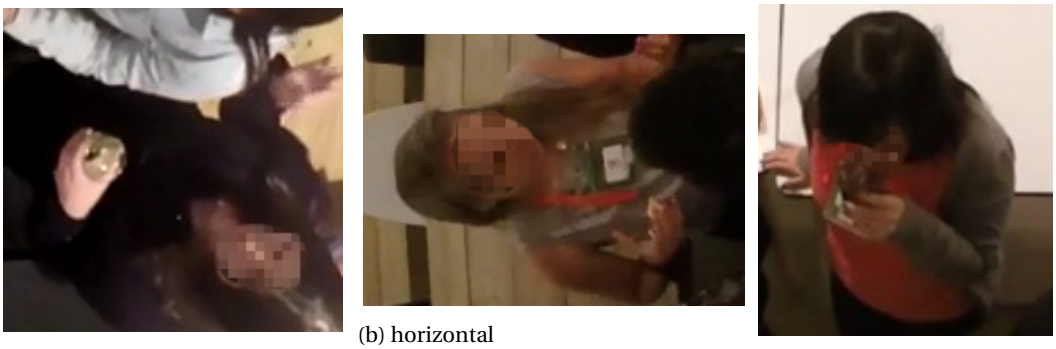
(a) static (b) drinking (c) moving & no-drinking



(d) static(DT) (e) drinking(DT) (f) moving(DT) & no-drinking

Figure 3.3: plots of multiple statuses of a subject with interesting points (red dots) and their corresponding dense trajectories(green curves)

orientation of subjects in frames. Also, the motion style is distinctive. These include motion intensity and motion frequency. Figure 3.4 gives some examples respect to the orientation of subjects.



(a) corner (b) horizontal (c) vertical

Figure 3.4: indication of subjects with a different orientation

3.3.3. Feature extraction

The second proposal is to extract the hog features from each bounding box and take it as the feature used for apparent description.

3.4. Fusion

The method of fusing both dynamic and static part given in this section. Normally, fusion is done by training a classifier on the concatenated scores which come from different classifiers. However, we fuse the results from different parts manually by using the motion level of the frame or window to calculate the contribution weight belongs to each of them. This is based on our assumption mentioned in Section 1.7.

3.4.1. Model

Figure 3.5 gives an overview about the fusion model. It could be divided into two parts. The dynamic part collects information brought by motion using Dense Trajectories(DT) while the static part catches information from the appearance of a static image or bounding box by using Bag of Features(BoF). Each section will have a score on how sure the current window is a drinking. If both of them agree on a same decision, then the classification process is finished. Otherwise, a ratio consists of both global motion weight and local motion weight will be calculated. According to this ratio, the scores from both sections are fused to make the final decision.

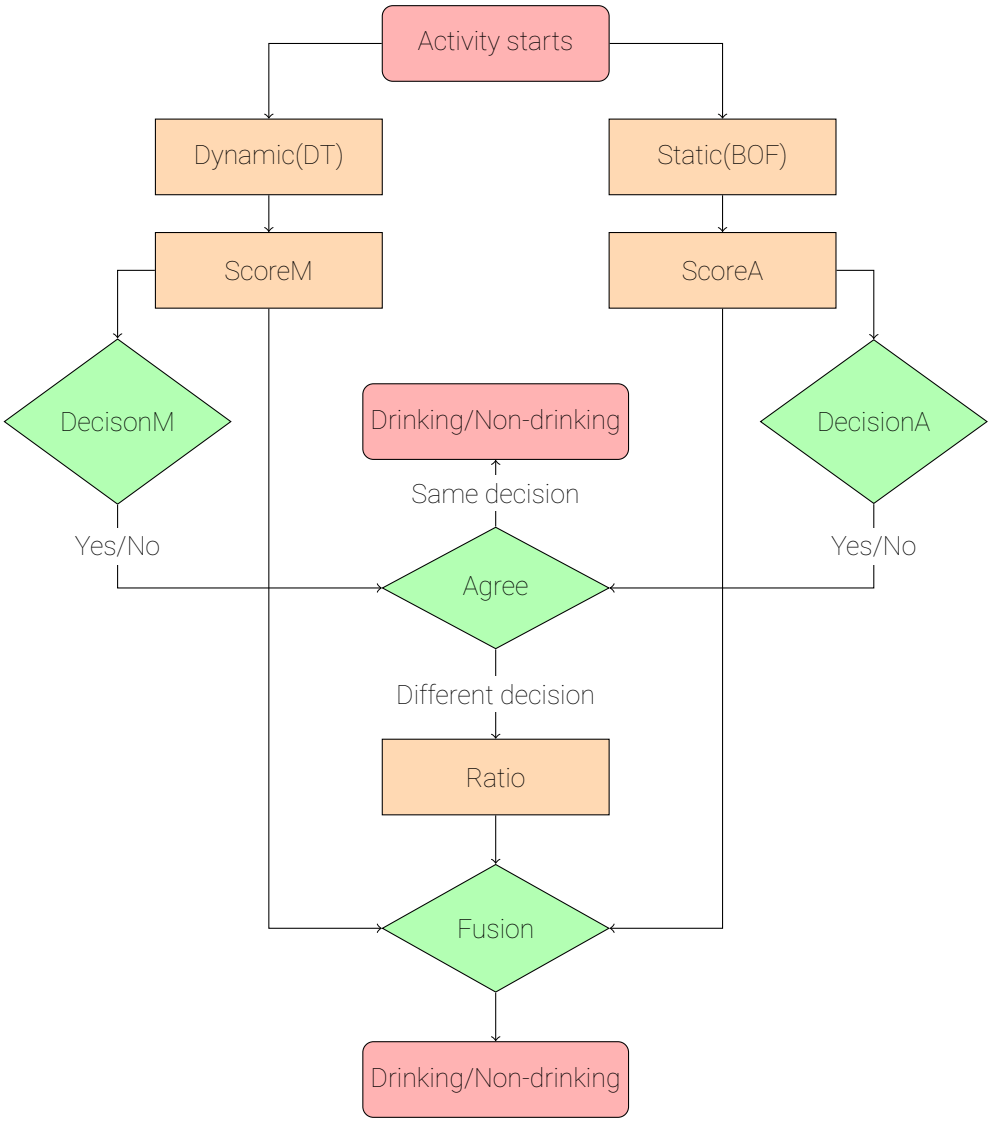


Figure 3.5: Drinking activity detection Flowchart

3.4.2. Fusion ratio

In action detection, both dynamic and static information are important. According to the fusion ratio, the dynamic and static scores for a window could be joined together to make a final decision. It contains two parts that are `global ratio` and `local ratio`. the global ratio respects to the whole duration (30 mins) and the local ratio respects to a single window (60 frames).

Global ratio

The function of the `global ratio` is to evaluate the motion level of a window over the entire period. Trajectory length is definitely the most crucial element in this section. Besides, the coordinate's variance in both x

and y direction are also important. Here, the motion feature vector is formed by 4 elements. They are listed as follows:

- m_l, σ_l^2 : the mean and variance of the trajectories' length in a window
- σ_x^2 : the variance of x coordinates of trajectories in a window
- σ_y^2 : the variance of y coordinates of trajectories in a window

The following paragraphs are used to describe how to get those elements from a set of trajectories in a window. For instance, if there are n trajectories $t_i, i = 1,2,...,n$ in a window(e.g. n = 30). for each t_i, a motion related feature vector could be formed as

$$x_i = [\sigma_{xi}^2, \sigma_{yi}^2, l_i, s_i]^T \tag{3.1}$$

where $l, \sigma_x^2, \sigma_y^2$ are the trajectory length, and variance in both x and y coordinates of the trajectory respectively. Besides, s_i describes the current trajectory is traced under which scale. All of them are defined in the section 3.2.1.

In Wang et al. [47], 8 spacial scales spaced by a factor of $1/\sqrt{2}$ are used. The global motion should be computed while a same scale of trajectory is used. Here, only the trajectories under the first scale are used for window global motion calculation. The set of selected feature vectors is given as $\{\mathbf{x}_i : s_i = 1\}$

Now, each elements in the global motion feature vector could be calculated as :

$$m_l = \frac{1}{n}\sum_{i=1}^{n} l_i \quad (s_i = 1) \tag{3.2}$$

$$\sigma_l^2 = \frac{1}{n}\sum_{i=1}^{n} (l_i - \mu_l)^2 \quad (s_i = 1) \tag{3.3}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_{xi}^2 \quad (s_i = 1) \tag{3.4}$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} \sigma_{yi}^2 \quad (s_i = 1) \tag{3.5}$$

As the target of this project is not finding the motion level of a subject in a window. A more precise way for motion level definition is to learn the weights for each element in the motion feature vector. Here, we only use the average of the trajectories' length ($\mathbf{m}_l$) for the global ratio calculation.

The global ratio of the current window i is calculated as

$$G_i = \frac{m_{li} - min(\mathbf{m}_l)}{max(\mathbf{m}_l) - min(\mathbf{m}_l)} \tag{3.6}$$

where $\mathbf{m}_l = (m_{l1}, m_{l2}, ..., m_{lk})$.

Local ratio

The ratio between the motion of the trajectories selected by MILES [8] and used for decision making and the motion of the rest trajectories in this window is called `local ratio`. The contribution of each instance $\mathbf{x}_{ij^*}$ to the classification of the window $\mathbf{W}_i$ is determined as

$$g(\mathbf{x}_{ij^*}) = \sum_{k \in I_{j^*}} \frac{\omega_k^* \mathbf{s}(\mathbf{x}^k, \mathbf{x}_{ij^*})}{m_k} \tag{3.7}$$

If $g(\mathbf{x}_{ij^*}) > \tau$, $\mathbf{x}_{ij^*}$ belongs to positive class; otherwise, $\mathbf{x}_{ij^*}$ belongs to the negative class. The parameter τ is chosen to be the same as $-\frac{b^*}{|U|}$ [8].

where the index set U is defined as

$$U = \left\{ j^* : j^* = arg\max_j exp\left(-\frac{||\mathbf{x}_{ij} - \mathbf{x}^k||^2}{\sigma^2} \right), k \in I \right\} \tag{3.8}$$

Now, we have the weights for both instances in the positive and negative class of window $\mathbf{W}_i$. And the `local ratio` is calculated as

$$L_i = \frac{\sum_{j=1}^{p} g(\mathbf{x}_{ij}) * l_{ij}}{\sum_{k=1}^{n} g(\mathbf{x}_{ik}) * l_{ik}} \tag{3.9}$$

where l_{ij} and l_{ik} are the normalized value of the clustered trajectory length in window i. The final fused ratio for the current window $\mathbf{W}_i$ could be written as

$$\beta_i = G_i * L_i \tag{3.10}$$

Fusion: weighted sum of the vectors vote

Figure 3.5 shows there are three inputs of the decision block `fusion`. They are `ScroreM`, `ScoreA` and `Ratio` respectively. Assume that the ratio $\mathbf{R}_i$ has a value r, to make summation of the contribution from both dynamic and static part to 1, it should have

$$\sum_{i=1}^{n} ScoreA_i = \beta_i$$
$$\sum_{j=1}^{m} ScoreM_j = 1 - \beta_i$$

The Error for the current window is defined as

$$E_i = (1-\beta_i) * \sum_{\theta=1}^{n} \alpha_\theta e^{-y_i \gamma_\theta k_s(x_\theta)} + \beta_i * e^{-y_i \sigma_i k_m(T_i)}, \quad y_i \in \{1,-1\} \tag{3.11}$$

where β_i is the fusion ratio of the current window. Here, I also introduce the notation $k_s(x_\theta)$ which denotes the predicted label of the current bounding box x_θ and the notation $k_m(T_i)$ is the predicted label of the current window using the clustered trajectories. Parameters α_θ describe the influence of the decision in frame θ on the overall decision of the current window if only static information used. The simplest choice for α_θ is

$$\alpha_\theta = \frac{1}{n} \tag{3.12}$$

while can be interpreted as decisions from all previous frames contributing equally to the final decision. This is a good choice when we consider all detection of the action equally valuable, regardless of *when* they are obtained. Whether all detection are considered equally valuable will depend on the nature of the problem-for instance, in the case of the hand moving towards the lips, the later detection would probably be more valuable, as they would have less motion information to use, also, the appearance looks more like a drinking. One possible way of giving more weight to the newer detection is making the weight negatively correlated with the motion level which is defined as the average motion over the previous 20 frames. In this case, it can be shown that the parameters α_θ are:

$$\alpha_\theta = \frac{m_{max} - m_\theta}{m_{max} - m_{min}} \tag{3.13}$$

At the end, the ratio between E_p and E_n is used as the score for current window. E_p is calculated when y_i is 1 and E_n is calculated when y_i is -1. The smaller the ratio is, the higher the confidence of the current window to be positive is. `E` stands for the predicted error.

4

Experiment Setup

In this section, the configuration of each experiment is explained. It includes the initial dataset formulation for both training and testing. Also, the motivation of each experiment is given. For experiments, two types of configurations are involved in this project. The first one is called `Subject Specific` (SS) and the second one is entitled as `Subject Independent`(SI). In SS, both training and testing dataset come from the same participant while SI may have them from different participants.

The experiments could be divided into three categories, one is relating to subject specific and the other one is relating to subject independent. Besides, there is another division under these two of categories. So for both categories, a set of motion related and static related experiment are set separately.

4.1. Subject selection

The first thing that needs to be clarified is the dataset using in all experiments. The MatchNMingle [4] dataset contains 92 participants. However, some of them have a high loss rate (over 40%). Moreover, some other participants have only taken a few drinking actions during the whole period. These will bring unexpected effects on the result of the experiments. Therefore, a group of subjects are removed from the original dataset before the actual experiments to avoid it. They are listed as follow:

- Subjects have a loss rate higher than 40% . Here, the loss rate stands for the time that a subject is out of the scene (all 3 cameras) or no ground truth available. [7, 11, 19, 44, 48]
- Subjects have the number of positive bags lower than 10. [7, 10, 17, 19, 29, 37, 41, 42, 44, 50, 51, 55, 65, 77]

So, in total, there are 76 subjects taken as the dataset for experiments related to subject specific. To remain the consistence, the experiments of subject independent also use the same dataset.

4.2. The appearance-based features

SURF features are extracted from the selected features point locations. The GridStep is [8 8] and the BlockWidth is [32 64 96 128]. K-means clustering is used to create a 500(default) word visual vocabulary.

4.3. Subject Specific

The nested cross validation is used for both parameter tuning and model validation. the level for tuning with 5 folds while the level for validation has 10 folds. So each fold has at least 1 positive bag as we keep only the subjects with at least 10 positive bags for experiments.

4.4. Subject Independent

In total, there are 2258 positive bags and 42049 negative bags. This is too large for calculation, the creation of the matrix of distance **m(B)** for the MILES becomes significantly consuming in terms of memory and computing time. So we first divide the whole data into 10 folds in a stratified manner. therefore, each fold will have around 240 positive bags and 4200 negative bags. 0.6203 (l: 0.0005, KPAR: 20)

4.4.1. Parameter tuning

In the MILES model, there are two parameters need to be tuned. They are `l` and `KPAR` separately. In order to tune the value of parameters in a reasonable range, we have to understand what is used for in the model. In the model, the importance of an instance on the final decision making is evaluated as a weight. the value of weight is ranged from 0 to 1. `KPAR` is used for defining this range. Another parameter `l` is used for setting a threshold. this threshold decides how many instances will be taken for the final decision making. The larger the value of `l` is, the less of instances will be used for the final decision making.

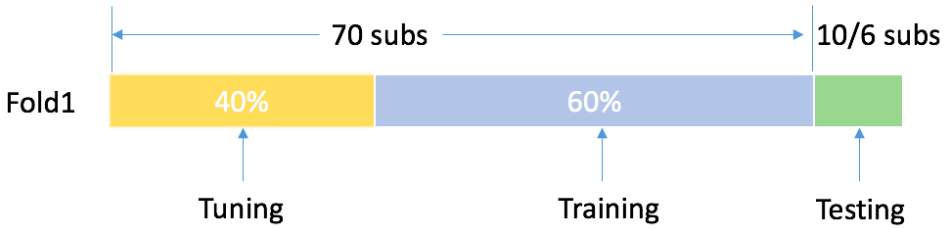


Figure 4.1: Dataset division scheme in Subject Independent case, here only gives 1 out of 11 folds as an example

The tuned range for these parameters in each case are given as follows:

- Subject specific
  - `l` : the range is from 0.001 to 0.1 with a step size 0.0005
  - `KPAR`: the range is from 15 to 25 with a step size 1
- Subject independent
  - `l` : the range is from 0.0001 to 0.1 with a step size 0.0005
  - `KPAR`: the range is from 15 to 25 with a step size 1

The range for `KPAR` for both cases is the same while the minimum value of `l` in subject independent case has a smaller value. This is because a much larger amount of instances are used in the second case, hence, more instances needed for the decision making.

4.5. Measures

Measurements are used to assess what is the performance of an architecture. However, before a measure process, it has to be clear that what are the metrics could be used for performance evaluation. [50] proposed some other standards besides of the classical one such as Recall, and Precision, as well as ROC curve.

- Logarithmic Loss Logarithmic Loss or Log Loss, works by penalizing the false classifications. It works well for multi-class classification. When working with Log Loss, the classifier must assign probability to each class for all the samples. Suppose, there are N samples belonging to M classes, then the Log Loss is calcualted as below:

$$Logarithmic\ Loss = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} * log(p_{ij}) \tag{4.1}$$

  where,
  y_ij, indicates whether sample i belongs to class j or not
  p_ij, indicates the probability of sample i belonging to class

5

Experiment Result

5.1. Measures of imitation

After the definition of the metrics could be used to describe a drinking behavior performed by our participants, the next step is to define when an imitation is occurred. An often-used experimental paradigm to test these imitation processes consists of exposing participants to condition in which confederates vary in the levels and pace of their alcohol consumption. By comparing individual's consumption across such conditions, it is possible to stringently test whether people adapt their drinking behavior to that of others. Nonetheless, we do not have this confederates participants set in the MatchNMingle dataset. Some other standard is needed. Luckily, we have group formulation information, A hypothesis that imitation will only happen between participants in the same group, therefore, the comparison is only necessary for ingroup participants. Here, we implemented a method mentioned in [28]. Imitation is scored as '1' if a a participants' sip taken within 10s after another group members' sip. No imitation is scored as '0'. In addition, the proportion of participants' previously imitated sips before their imitated sip (i.e., number of previously imitated sips divided by the number of the group members' previous sips) is also examined.

5.2. Subject specific

In this section, subjects are investigated individually.For `auc` calculation, both training and test data come from the same subject in a two level nested cross-validation. The first level is used for auc calculation with 10 folds and the second level is used for parameter tuning with 5 folds. To make sure each fold has at least 1 positive bag, the subjects with a positive bag number lower than 10 are removed from the data using for experiments.

5.2.1. Motion-based

Figure 5.1 gives the result with only motion cues used. The red dot line represents the average auc that has a value of 0.62. The range at the top of each blue bar are the variance introduced by cross validation.

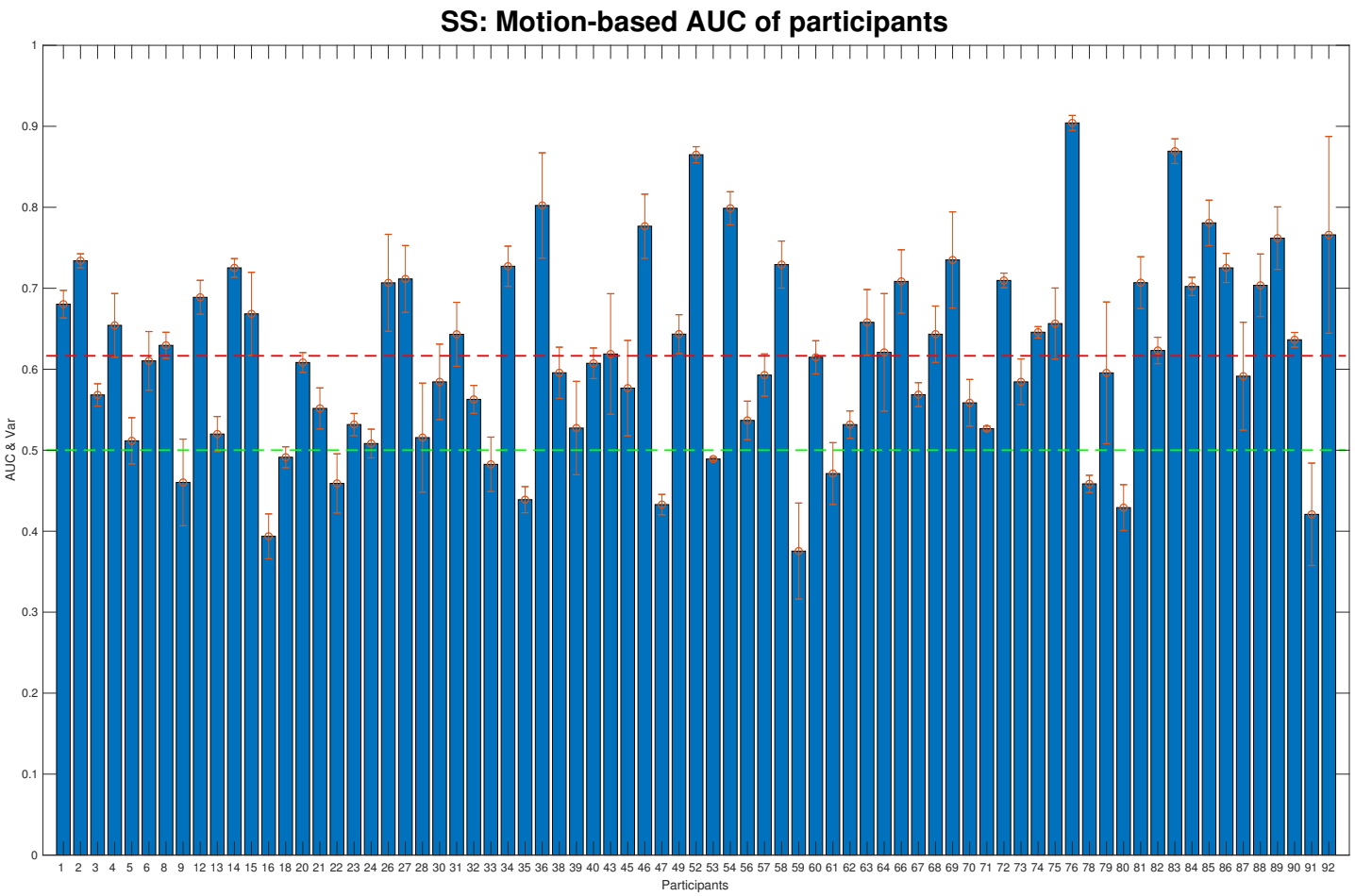


Figure 5.1: `auc` and `var` calculated with normalized features used for each participant while a nested cross-validation is used for model parameter turning. The `mean` of the `auc` over all participants is 0.7721 indicated by the red dot line while the green dot line shows the value of 0.5 located.

5.2.2. Appearance-based

Figure 5.2 gives a comparison result between the calculated posterior probability and the ground truth. In this experiment, all bounding boxes of subject 3 were assigned into 10 folds in a stratified way. one out of them is used for testing and the rest for training. In the figure, color blue represents for ground truth while the orange line are calculated posterior probability of each testing bounding box while only static information is used. The probability lower than 0.5 are suppressed to 0. If most of the PP (posterior probability) in a blue bar have a value higher than 0.5. Based on that definition, eight out of nine drinking actions are detected correctly. However, there are more than 4 positions were detected as drinking while it is actually not (orange peak is not in a blue bar).

The bounding box with the coordinates (2424,1) is given as figure 5.3. The appearance shows that the subject has a high probability of drinking. However, the ground truth annotated this bounding box as not. In this situation, The introduce of motion section is assumed to reduce this false positive rate.

Figure 5.4 gives the result while only appearance information used. the AUC calculated for all 76 participants. The error bar on the top of each stem stands for the variance calculated in the cross-validation process. The red dot line is the mean value (0.8983) of all measured results.

Table 5.1: The detail result for the test set of fold1, subject3

		Predicted	
		Drinking	No-Drinking
Actual	Drinking	0.72	0.28
	No-Drinking	0.18	0.82

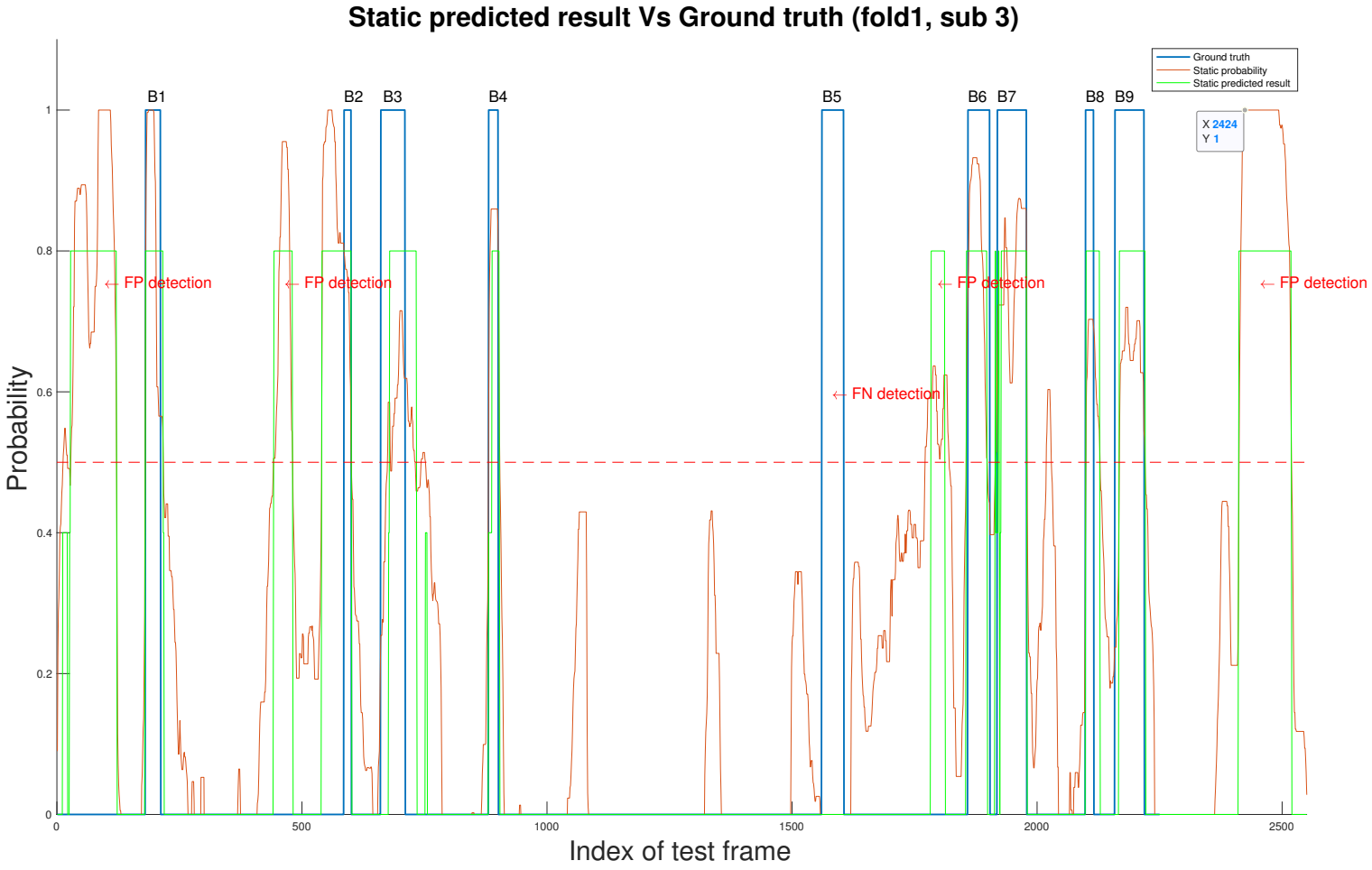


Figure 5.2: A comparison between appearance-based predicted result(green line) and the ground truth(blue line) for fold 1, sub 3

5.2.3. Hybrid

The key element in this testing process is how many positive bags that we can have for training. It is also true for these same type problems. Therefore, we decided to use half of the positive samples for training and the rest used for testing. To avoid the overlap between the training and testing dataset. one tenth video segment of the subject is used as the gap between them. In some cases, the total number of instances for a single subject is too large for MATLAB to generate the matrix `m`. So a simple sampling process which selects the same negative bags as the positive one is used to decrease the dimension of `m`.

Figure 5.3: The corresponding frame with a high score calculated by the appearance-based classifier

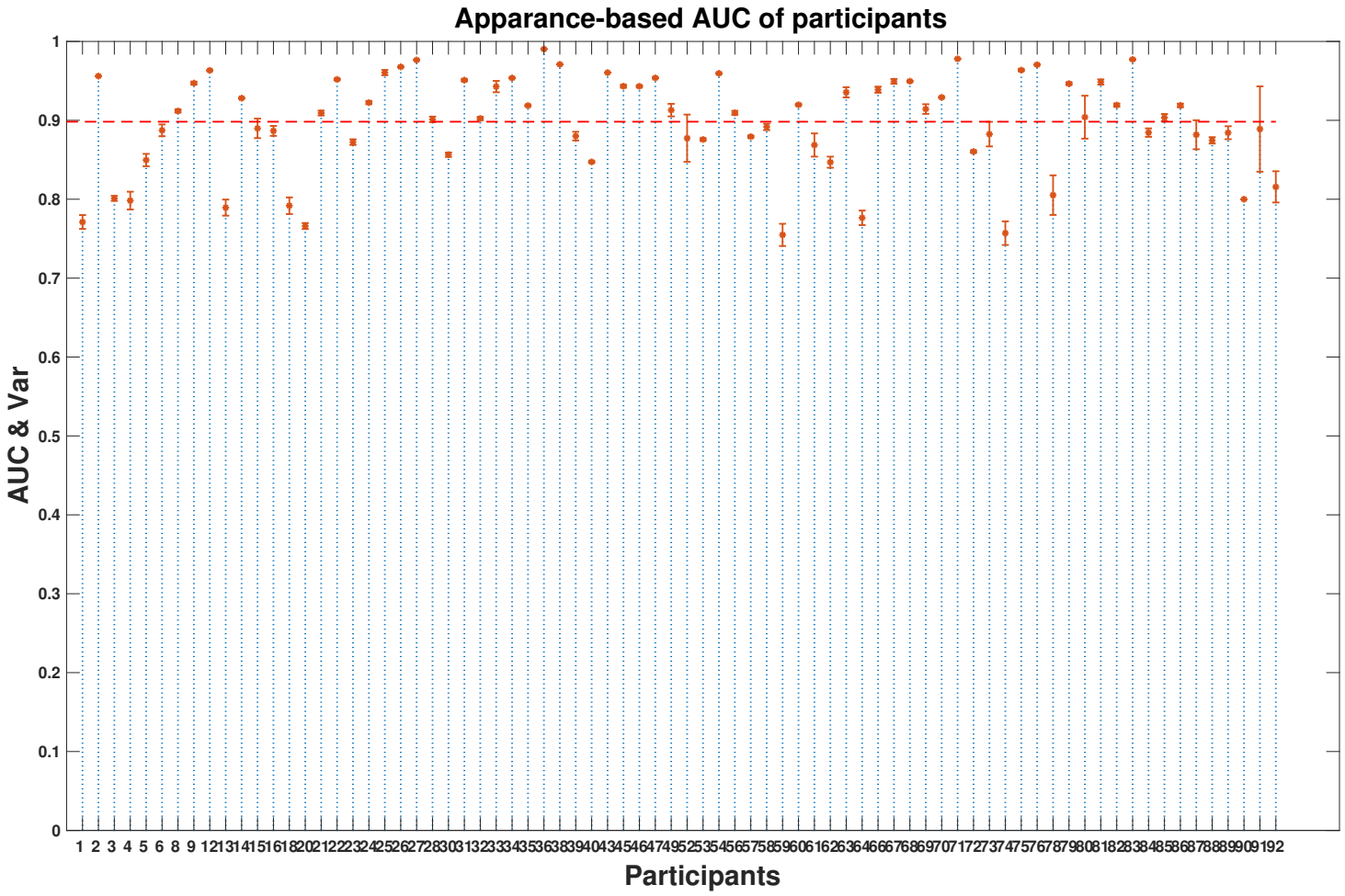


Figure 5.4: AUC of participants while only shape information used for subject specific

5.2.4. Result

Table 5.2 shows the result of experiment for SS. The best performance is when only appearance information is used for action detection. The motion-based method gives the worst performance.

Table 5.2: Results of threes scenarios in Subject Specific case

Subject Specific Result	
Method	AUC
Appearance	0.89±0.014
Motion	0.62±0.014
Fusion	0.73 ± 0.025

5.3. Subject independent

Leave-n-subjects-out cross validation Tabel 5.3 shows the result of the experiments for SI. Leave n subjects out is used as the cross validation. The baseline is the result shown in [5] while a general gesture detection is studied. In general, our result of drinking detection is not as good as the baseline. Overall, fusion presents the best result in our case.

Table 5.3: Results of three scenarios in Subject Independent case

Subject Independent Result	
Method	AUC
Appearance	0.52± 0.045
Motion	0.48± 0.014
Fusion	0.59± 0.032

6 Discussion

6.1. Threshold selection

Frame-level classifier is computationally expensive. Hence, a threshold is used to filter the frames with a relatively high motion level out. We have assumed the appearance information plays a less important role while the motion level of frames are relatively high. It helps us to reduce the computational requirement. So, what should the value of the threshold be. The first try is to set it at the mean of the motion level of all drinking frames. The result is shown in Table 6.2. The True Positive rate decrease from 0.48 to 0.44 while we have increased the threshold from 9 to 10. Of course, we can not decide on which value that we should use for this threshold only based on two experiments. We leave it as the future work due to the tuning process is fairly time consuming. In this project, we select 9 as the threshold.

Table 6.1: The matrix for the first configuration. The training set consists of 14827 images while the threshold is set to 9

		Predicted	
		Drinking	No-Drinking
Actual	Drinking	0.48	0.52
	No-Drinking	0.44	0.56

Table 6.2: The matrix for the second configuration. The training set consists of 36206 images while the threshold is set to 10

		Predicted	
		Drinking	No-Drinking
Actual	Drinking	0.44	0.56
	No-Drinking	0.39	0.61

6.2. Validation method selection

The parameter tunning on motion-based classifier plays an essential role in this project. It is specially because of the parameters used in MILES. So the tuning strategy combined with the validation method is another topic should be discussed. Two totally different situations appear in SS and SI. In SS, the dataset that can be used for training, testing and validation is limited. We have to use the data as efficient as passive. Therefore, a two level cross-validation is implemented. The first level of cross validation with 5 folds is used for parameter tuning while the second level of cross validation with 10 folds is used to evaluate the model. Each fold is ensured to have at least one positive bag as only the subjects with at least have 10 positive bags are sued in experiments. For SI, the whole dataset with 76 subjects are divided directly into 3 groups because the dataset is large enough. Leave-n-out cross validation is used here. 76 subjects are divided into 11 groups. So each fold has either 6 or 10 subjects for testing and the rest is used for training. As shown in 4.1. the rest 70 subjects are divided into other 2 groups with 40% and 60 % of this 70 subjects in a stratified way. 40 % is used for parameter tuning while the other 70 % is used for traing in the current fold.

6.3. Ground truth vs Window labels

To compare the fused result with the individual one or compare results between individuals, such as static and dynamic, we have to standard the metrics that we will use for evaluating all results. Here, we use AUC respecting to a frame instead of a window. In paper [5], the calculated AUC respects to a window. However, as shown in Figure 6.1, the error has already introduced if a majority voting used for determining the labels of windows. Bar 2, 4 and 8 are regarded as negative bags, although the ground truth says they are drinking. Hence, it is not a good idea of taking the AUC respect to a window as the metric for the tested model evaluation in this case.

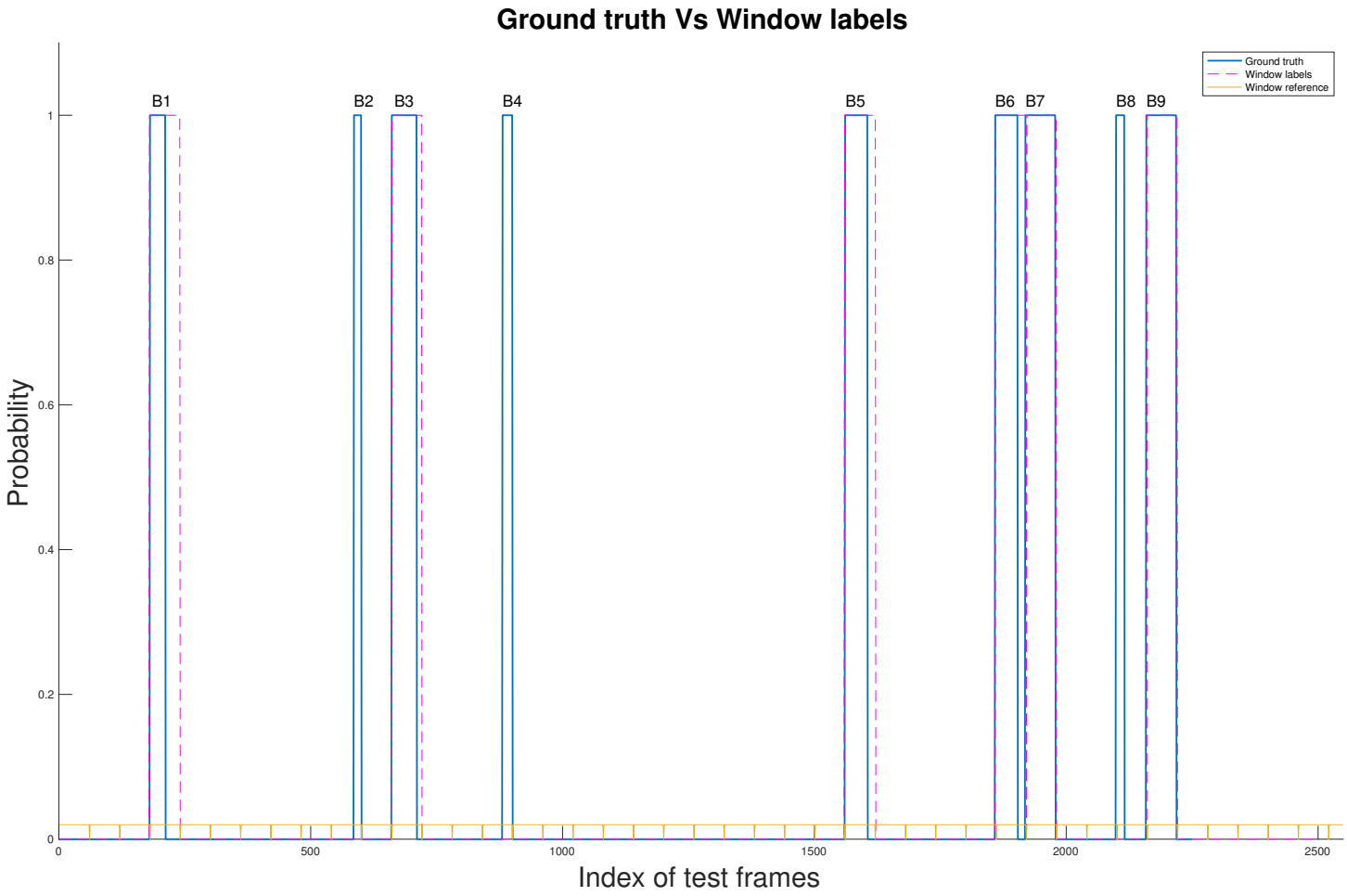


Figure 6.1: A comparison between the determined window labels with majority voting utilized and the ground truth for fold1, subject3

6.4. Number of positive bags vs AUC

In this section, the relation between the quantity of positive bags and the calculated AUC for each participant is examined. In Figure 6.2, the number of positive bags is represented as blue bars and the calculated AUC are noted as red crosses on top of them. The value of AUC is multiplied by 100 times to put the two factors on the same scale. In order to find the relation, the number of bags is plotted in ascending order. Additionally, the green curve fitted from all AUC values provided. We can see that there is no explicit connection could be observed between them. A slight drop is seen on the variance of AUC along with the growth in the number of positive bags. The highest AUC is achieved while the number of positive bags is about 28. Moreover, the red and blue dot line denotes the mean of AUC (0.77) and the number of bags(30) respectively.

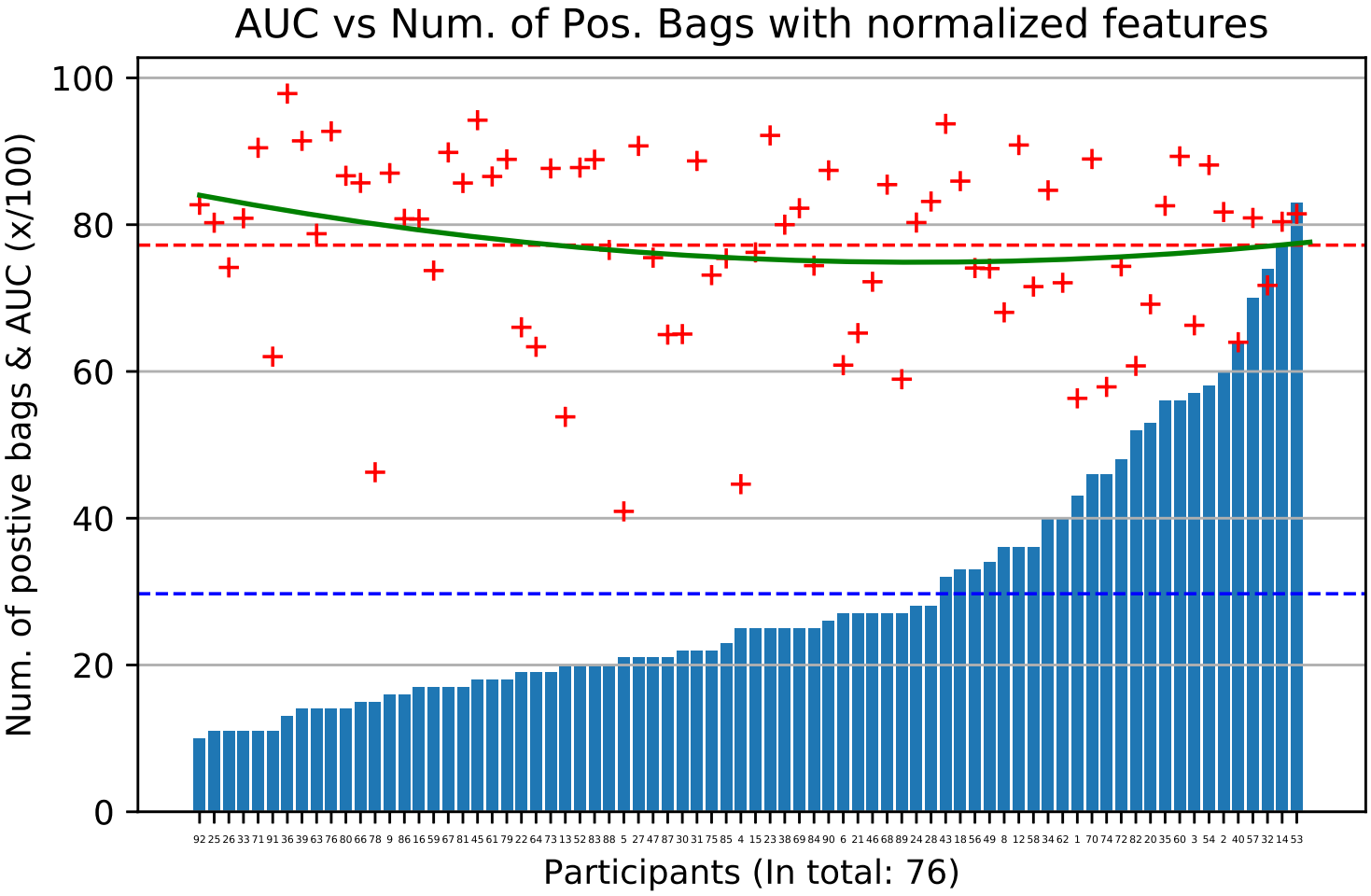


Figure 6.2: AUC calculated with normalized features used for each participant while nested cross-validation is used for model parameter turning. The `mean` of the AUC over all participants is 0.7721 indicated by the red dot line. In addition, the bar graph shows the `number of positive bags` each participant has while the blue line shows the `average` bag number of all participants

7 Conclusion

The aim of this project is to localize the drinking action in the temporal domain in crowded mingle scenarios. We propose to use both appearance and motion features to achieve this target. The action detection tasks in MatchNMingle have additional challenges due to the videos were recorded in a top view. It leads to the result that the extracted action trajectories emitted by different participants are not consistent both in length and direction. For instance, people located at the left part of the scene (shown in Fig. 1.1) have the trajectories(used to describe drinking action) moving from right to left and people stand at the right side have the trajectories with the opposite direction. BoF and MILES are implied to overcome this variety.

SS and SI are the two types of experimental configurations studied in this report using our proposed model. As expected, the performance of experiments that both training and testing dataset come from the same subject is higher than the case that a different set of subjects are used for training. In SS, the appearance-based classifier gives the best performance while it appears in the fusion model in SI. In addition, we found the temporal splitting with fixed window size and labelling them by using majority voting may introduce some fundamental error in the labeling stage. Another methodology splitting a dataset with a different window size is elaborated in Sec. 7.1. To avoid the bias introduced by the labeling stage, we take AUC in frame-level as the metric to evaluate our model instead of using it with respecting to a window.

Furthermore, we investigated the relationship between the number of positive bags and the calculated AUC with the motion-based classifier. there is no evidence shown that a correlation relation exists between them.

Besides, we have proved that static information plays a crucial role in the drinking action detection besides motion cues. Also, a fusion model that can use both results from two independent classifiers are defined. The weight given to each segment is calculated based on the motion level of a window. Nevertheless, the later experiments indicate that the fusion of information respect to a window is redundant. We can simply fuse the information in frame-level as the AUC is calculated respect to a frame. Although the fusion model does not give an obvious improvement over individual models, there is also no proofs saying fusion is not the right direction to go. More parameter tuning work may need for the model to be able to have a higher performance.

7.1. Future Work

This project conducts drinking action detection under the assumption that memory and time are unlimited. However, this is not true in real life. In practice, personal behavior detection is most likely to be applied in a real-time situation. Then, the computational efficiency might be a valuable point for future researches especially when the frame-level experiments are involved as they are computationally expensive both in memory and time.

Another direction is to apply DT in an orientation irrelevant way. The results of experiments from SS give a much better performance compared with the results from SI while only the motion information is applied. One of the guesses is the drinking action related trajectories appearing in a different orientation from the camera view. A possible solution that is analogous to SIFT could be rotating the exist trajectory in a few directions and consider the set of all rotated and the original trajectories as a unit. In MILES, The similarity is calculated between each unit instead of a single trajectory.

Furthermore, 3D space-time descriptors could be used for drinking action description due to its big success in general human action recognition and localization. Here, we have only used 2D descriptors extracted respect to a single frame and then aggregating their decisions made on each frame with a manually calculated weight. However, a 3D descriptor which can extract features from multiple frames is worth to try in this case. Some other potential topics on drinking action detection in crowded situation are discussed as follow.

7.1.1. Action localization

The spatial extent of the action is already fixed due to ROI in MatchNMingle is given. Therefore, we only need to delimit the beginning and length of an action. In this work, the temporal slices used in both training and testing dataset with a fixed length have introduced some problems. Error is introduced when we label the sliding window with majority voting. It is shown in Fig. 6.1. Kläser et al. [20] have proposed a more advanced way to do this job. When training the sliding window classifier, the temporal slices are aligned with the ground-truth begin and end time stamps of the action. At test time, a sliding window with multiple temporal scales is used to localize actions.

7.1.2. Samples selection

In MatchNMingle, the number of positive samples is much less than the number of negative one in drinking action detection. How to select the most representative samples turns to be one of the key questions in this problem. Samples selection consists of two segments. They are motion descriptor and image descriptor selection respectively. It is needed as the video related analyzes are computational heavily, hence, only the descriptor with the highest discriminative power are used as the input of the model. In the current stage, samples of motion are randomly selected in a stratified way while the static features are simply filtered out by setting a threshold in an intuitive way.

7.1.3. Fusion model

In this work, we present the fusion method that gives weights to motion and static based on the motion level of the current frame and the overall motion level of the current window. All scores of both parts are manually integrated together. There may be a way that can learn the weights automatically from the concatenated scores calculated by both classifiers.

Acknowledgements

I would like to express my special thanks of gratitude to my supervisor Assoc. Prof. dr.ir.H.Huang as well as my daily supervisor Dr. L. Cabrera-Quiros who gave me the golden opportunity to do this wonderful project on the topic drinking behavior detection in a crowded setting, which also helped me in doing a lot of research and I came to know about so many new things I am really thankful to them. Secondly, I would also like to thank my parents and girlfriend who helped me a lot in finalizing this project within the limited time frame.

Bibliography

[1] Timur Bagautdinov, Alexandre Alahi, François Fleuret, Pascal Fua, and Silvio Savarese. Social scene understanding: End-to-end multi-person action localization and collective activity recognition. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-January, pages 3425–3434. Institute of Electrical and Electronics Engineers Inc., 11 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.365.

[2] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea Cavallaro. Video-based human behavior understanding: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(11):1993–2008, 2013. ISSN 10518215. doi: 10.1109/TCSVT.2013.2270402.

[3] Molly A. Bowdring and Michael A. Sayette. Perception of physical attractiveness when consuming and not consuming alcohol: a meta-analysis. *Addiction*, 113(9):1585–1597, 2018. ISSN 13600443. doi: 10.1111/add.14227.

[4] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij, and Hayley Hung. The MatchNMingle dataset: a novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates, 2018. ISSN 19493045. URL `https://ieeexplore.ieee.org/document/8395003/`.

[5] Laura Cabrera-Quiros, David M.J. Tax, and Hayley Hung. Gestures in-the-wild : detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. pages 1–10, 2018.

[6] Barry D. Caudill and Fan Hui Kong. Social approval and facilitation in predicting modeling effects in alcohol consumption. *Journal of Substance Abuse*, 13(4):425–441, 2001. ISSN 08993289. doi: 10.1016/S0899-3289(01)00099-2.

[7] Rama Chellappa, Amit K. Roy-Chowdhury, and S. Kevin Zhou. Recognition of Humans and Their Activities Using Video. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2005. ISSN 1559-8136. doi: 10.2200/S00002ED1V01Y200508IVM001.

[8] Y Chen, J Bi, and J Wang. {MILES}: multiple instance learning via embedded instance selection. *{IEEE} Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1–17, 2006.

[9] Guilhem Chéron, Anton Osokin, Ivan Laptev, and Cordelia Schmid. Modeling Spatio-Temporal Human Track Structure for Action Localization. 6 2018. URL `http://arxiv.org/abs/1806.11008`.

[10] M. L. Cooper. Motivations for alcohol use among adolescents: Development and validation of a four-factor model. *Psychological Assessment*, 6(2):117–128, 1994. ISSN 10403590. doi: 10.1037/1040-3590.6.2.117.

[11] Navneet Dalal and Bill Triggs. Histograms of Oriented Gradients for Human Detection. Technical report. URL `http://lear.inrialpes.fr`.

[12] Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: a study of bag-of-features and part-based representations. *Procedings of the British Machine Vision Conference 2010*, pages 1–97, 2010. doi: 10.5244/C.24.97. URL `https://www.di.ens.fr/~josef/publications/delaitre10.pdfhttp://www.bmva.org/bmvc/2010/conference/paper97/index.html%5Cnhttp://action-classif.googlecode.com/svn-history/r29/trunk/slidesOxford/slides.pdf%5Cnhttp://www.bmva.org/bmvc/2010/conference/paper97/`.

[13] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior Recognition via Sparse Spatio-Temporal Features. *Journal of Clinical Medicine*, 7(3):52, 2005. doi: 10.3390/jcm7030052. URL `http://vision.ucsd.edu`.

[14] Korrina A. Duffy and Tanya L. Chartrand. Mimicry: Causes and consequences. *Current Opinion in Behavioral Sciences*, 3:112–116, 2015. ISSN 23521546. doi: 10.1016/j.cobeha.2015.03.002. URL http://dx.doi.org/10.1016/j.cobeha.2015.03.002.

[15] Yoav Freund and Robert E Schapire. Journal of Computer and System Sciences s SS1504 journal of computer and system sciences. Technical report, 1997. URL https://ac.els-cdn.com/S002200009791504X/1-s2.0-S002200009791504X-main.pdf?_tid=3fb4167e-9c0f-403f-8a17-2a9a19f3a499&acdnat=1537711548_9213d0f23b140ed2fb8bf8b2cb4e2d93https://ac.els-cdn.com/S002200009791504X/1-s2.0-S002200009791504X-main.pdf?_tid=773174d.

[16] N Gueguen, C Jacob, and A Martin. Mimicry in social interaction: Its effect on human judgment and behavior. *European Journal of Social Sciences*, 8(2):253–260, 2009. URL http://eyethink.org/resources/papers/Gueguen-et-al..pdf.

[17] a. Hauptmann. Action recognition via local descriptors and holistic features. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 58–65, 2009. ISSN 2160-7508. doi: 10.1109/CVPRW.2009.5204255. URL http://www.cs.cmu.edu/~./mychen/publication/XingHuanCVPR4HB09.pdfhttp://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5204255.

[18] Mihir Jain, Jan C. van Gemert, and Cees G.M. M. Snoek. What do 15,000 object categories tell us about classifying and localizing actions? In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 46–55. IEEE, 6 2015. ISBN 9781467369640. doi: 10.1109/CVPR.2015.7298599. URL http://ieeexplore.ieee.org/document/7298599/.

[19] A. Klaeser, M. Marszalek, and C. Schmid. A Spatio-Temporal Descriptor Based on 3D-Gradients. pages 1–99. British Machine Vision Association and Society for Pattern Recognition, 2 2012. doi: 10.5244/c.22.99.

[20] Alexander Kläser, Marcin Marszałek, Cordelia Schmid, and Andrew Zisserman. Human focused action localization in video. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6553 LNCS, pages 219–233, 2012. doi: 10.1007/978-3-642-35749-7{_}17.

[21] Jessica L. Lakin, Valerie E. Jefferis, Clara Michelle Cheng, and Tanya L. Chartrand. the Chameleon Effect As Social Glue:Evidence for the Evolutionary Significance of Nonconscious Mimicry. *Journal of Nonverbal Behavior*, 27(3):145–161, 2003. ISSN 1573-3653. doi: 10.1023/A:1025389814290.

[22] Laptev and Lindeberg. Space-time interest points. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 432–439. IEEE, 2003. ISBN 0-7695-1950-4. doi: 10.1109/ICCV.2003.1238378. URL http://ieeexplore.ieee.org/document/1238378/.

[23] I. Laptev. Improvements of Object Detection Using Boosted Histograms. In *Procedings of the British Machine Vision Conference 2006*, pages 1–97. British Machine Vision Association, 2006. ISBN 1-901725-32-4. doi: 10.5244/C.20.97. URL http://www.bmva.org/bmvc/2006/papers/434.html.

[24] Ivan Laptev. On space-time interest points. In *International Journal of Computer Vision*, volume 64, pages 107–123, 2005. doi: 10.1007/s11263-005-1838-7.

[25] Ivan Laptev and P Patrick. Natural movie scene action recognition Retrieving actions in movies. 2007.

[26] Ivan Laptev and Patrick Perez. Retrieving actions in movies. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. ISBN 978-1-4244-1630-1. doi: 10.1109/ICCV.2007.4409105. URL http://ieeexplore.ieee.org/document/4409105/.

[27] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. Technical report. URL www.weeklyscript.com.

[28] Helle Larsen, Rutger C M E Engels, Pierre M. Souren, Isabela Granic, and Geertjan Overbeek. Peer influence in a micro-perspective: Imitation of alcoholic and non-alcoholic beverages. *Addictive Behaviors*, 35(1):49–52, 2010. ISSN 03064603. doi: 10.1016/j.addbeh.2009.08.002. URL http://dx.doi.org/10.1016/j.addbeh.2009.08.002.

[29] Weiyao Lin, Ming-Ting Sun, Radha Poovandran, and Zhengyou Zhang. Human activity recognition for video surveillance. *IEEE International Symposium on Circuits and Systems*, pages 2737–2740, 2008. ISSN 02714310. doi: 10.1109/ISCAS.2008.4542023.

[30] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, pages 1150–1157. IEEE, 1999. ISBN 0-7695-0164-8. doi: 10.1109/ICCV.1999.790410. URL http://ieeexplore.ieee.org/document/790410/.

[31] Bruce D Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. Technical report.

[32] Shugao Ma, Leonid Sigal, and Stan Sclaroff. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. pages 1942–1950. Institute of Electrical and Electronics Engineers (IEEE), 12 2016. doi: 10.1109/cvpr.2016.214.

[33] Pyry Matikainen, Martial Hebert, and Rahul Sukthankar. Trajectons: Action Recognition Through the Motion Analysis of Tracked Features. Technical report.

[34] Kenji MATSUI, Toru TAMAKI, Bisser RAYTCHEV, and Kazufumi KANEDA. Trajectory-Set Feature for Action Recognition. *IEICE Transactions on Information and Systems*, E100.D(8):1922–1924, 2017. ISSN 0916-8532. doi: 10.1587/transinf.2017EDL8049. URL https://www.jstage.jst.go.jp/article/transinf/E100.D/8/E100.D_2017EDL8049/_article.

[35] Ross Messing, Chris Pal, and Henry Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 104–111, 2009. ISBN 9781424444205. doi: 10.1109/ICCV.2009.5459154.

[36] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification. Technical report.

[37] Xiaojiang Peng and Cordelia Schmid. Multi-region Two-Stream R-CNN for Action Detection. pages 744–759. 2016. doi: 10.1007/978-3-319-46493-0{_}45. URL http://link.springer.com/10.1007/978-3-319-46493-0_45.

[38] Bernt Schiele and James L Crowley Gravir. Recognition without Correspondence using Multidimensional Receptive Field Histograms. Technical Report 1, 2000.

[39] Christian Schuldt, Laptev Barbara, and Se Stockholm. Recognizing Human Actions : A Local SVM Approach. *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 3: 32–36, 2004. doi: 10.1109/ICPR.2004.1334462.

[40] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. page 357. Association for Computing Machinery (ACM), 10 2007. doi: 10.1145/1291233.1291311.

[41] Bharat Singh, Tim K. Marks, Michael Jones, Oncel Tuzel, and Ming Shao. A Multi-stream Bi-directional Recurrent Neural Network for Fine-Grained Action Detection. pages 1961–1970. Institute of Electrical and Electronics Engineers (IEEE), 12 2016. doi: 10.1109/cvpr.2016.216.

[42] Renske Spijkerman, Helle Larsen, Frederick X. Gibbons, and Rutger C M E Engels. Students' drinker prototypes and alcohol use in a naturalistic setting. *Alcoholism: Clinical and Experimental Research*, 34 (1):64–71, 2010. ISSN 01456008. doi: 10.1111/j.1530-0277.2009.01067.x.

[43] Ju Sun, Xiao Wu, Shuicheng Yan, Loong Fah Cheong, Tat Seng Chua, and Jintao Li. Hierarchical spatio-temporal context modeling for action recognition. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, 2009. ISBN 9781424439935. doi: 10.1109/CVPRW.2009.5206721.

[44] Javid Ullah and Muhammad Arfan Jaffar. Object and motion cues based collaborative approach for human activity localization and recognition in unconstrained videos, 2017. ISSN 15737543. URL https://link.springer.com/content/pdf/10.1007%2Fs10586-017-0825-4.pdf.

[45] Muhammad Muneeb Ullah, Sobhan Naderi Parizi, and Ivan Laptev. Improving bag-of-features action recognition with non-local cues. In *Procedings of the British Machine Vision Conference 2010*, pages 1–95. British Machine Vision Association, 2010. ISBN 1-901725-40-5. doi: 10.5244/C.24.95. URL http://www.bmva.org/bmvc/2010/conference/paper95/index.html.

[46] Rick Van Baaren, Loes Janssen, Tanya L. Chartrand, and Ap Dijksterhuis. Where is the love? The social aspects of mimicry. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1528): 2381–2389, 2009. ISSN 14712970. doi: 10.1098/rstb.2009.0057.

[47] Heng Wang, Alexander Kl, Cordelia Schmid, Liu Cheng-lin, Heng Wang, Alexander Kl, Cordelia Schmid, Liu Cheng-lin Action, and Alexander Kl. Action Recognition by Dense Trajectories To cite this version :. 2011.

[48] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 103(1):60–79, 2013. ISSN 09205691. doi: 10.1007/s11263-012-0594-8.

[49] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5303 LNCS, pages 650–663, 2008. ISBN 3540886850. doi: 10.1007/978-3-540-88688-4-48. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.471.6845&rep=rep1&type=pdf.

[50] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, Christophe Garcia, and Bülent Sankur. Evaluation of video activity localizations integrating quality and quantity measurements q. *Computer Vision and Image Understanding*, 127:14–30, 2014. doi: 10.1016/j.cviu.2014.06.014. URL http://dx.doi.org/10.1016/j.cviu.2014.06.014.

[51] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 492–497, 2009. ISBN 9781424444205. doi: 10.1109/ICCV.2009.5459201.

[52] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. End-to-end Learning of Action Detection from Frame Glimpses in Videos. 11 2015. URL http://arxiv.org/abs/1511.06984.