AUTOMATED DETECTION OF VIEWER ENGAGEMENT BY HEAD MOTION ANALYSIS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS INSTITUTE
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

UĞUR GÜLER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

APRIL 2015

AUTOMATED DETECTION OF VIEWER ENGAGEMENT BY HEAD MOTION ANALYSIS

Submitted by **Uğur GÜLER** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems**, **Middle East Technical University** by,

Prof. Dr. Nazife Baykal ____________________
Director, **Informatics Institute, METU**

Prof. Dr. Yasemin Yardımcı Çetin ____________________
Head of Department, **Information Systems, METU**

Assoc. Prof. Dr. Alptekin Temizel ____________________
Supervisor, **Modeling and Simulation, METU**

Examining Committee Members:

Assoc. Prof. Dr. Altan Koçyiğit ____________________
Information Systems, METU

Assoc. Prof. Dr. Alptekin Temizel ____________________
Modeling and Simulation, METU

Assoc. Prof. Dr. Aysu Betin Can ____________________
Information Systems, METU

Assist. Prof. Dr. Erhan Eren ____________________
Information Systems, METU

Assoc. Prof. Dr. Hüseyin Hacıhabiboğlu ____________________
Modeling and Simulation, METU

Date: **17/04/2015**

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and result that are not original to this work.

Name, Last Name	**:**	**Uğur Güler**
Signature	**:**	

ABSTRACT

AUTOMATED DETECTION OF VIEWER ENGAGEMENT BY HEAD MOTION ANALYSIS

Güler, Uğur

M.S., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Alptekin Temizel

April 2015, 48 pages

Measuring viewer engagement plays a crucial role in education and entertainment. In this study we analyze head motions of the viewers from video streams to automatically determine their engagement level. Due to unavailability of a dataset for such an application, we have built our own dataset. By using face detection system, the head position of viewer is obtained throughout the video for each frame. Then, using these positions, we analyze and extract some features. In order to classify the data, we employ both Random Forest and Support Vector Machine (SVM) with extracted parameters. User engagement detection is performed using the employed model and the results indicate accuracy of 89.4% and recall of 90.9% on our dataset with Random Forest.

Keywords: Face Detection, Machine Learning, Video Content Analysis

ÖZ

İZLEYİCİ İLGİ SEVİYESİNİN KAFA HAREKETLERİNİN ANALİZİ İLE OTOMATİK TESPİTİ

Güler, Uğur

Yüksek Lisans, Bilişim Sistemleri Bölümü

Tez Yöneticisi: Doç. Dr. Alptekin Temizel

Nisan 2015, 48 sayfa

İzleyici ilgi seviyesinin tespiti eğitim ve eğlence alanlarında önemli rol oynar. Bu çalışmada video akışındaki kafa hareketlerini analiz ederek izleyicinin ilgi seviyesi otomatik olarak tespit etmek hedeflenmiştir. Bu uygulama için hâlihazırda bir veri seti olmadığından dolayı kendi veri setimizi oluşturduk. Yüz tespit sistemi kullanarak, videodaki izleyici kişinin video boyunca her bir kare için kafa pozisyonları elde edilmiştir. Elde edilen bu pozisyon verilerinden, kafa hareketleri analiz edilmiş ve bazı nitelikler çıkartılmıştır. Verileri sınıflandırabilmek için Rastgele Orman algoritması ve Destek Vektör Makineleri ayrı ayrı çıkartılan bu nitelikler ile çalıştırılmıştır. Model kullanılarak ilgi seviyesi ölçülmüş ve Rastgele Orman algoritması ile %89.4 doğruluk, %90.9 duyarlılık oranında sonuçlar elde edilmiştir.

Anahtar Sözcükler: Yüz Tanıma, Makine Öğrenimi, Video İçerik Analizi

To my dear wife

ACKNOWLEDGMENTS

I would like to thank my supervisor Assoc. Prof. Dr. Alptekin Temizel for his encouraging, advice and guidance in this thesis study.

I am also thankful to Ayşe Elvan Gündüz for her help and replied kindly whenever I had a question.

I would like to thank my family, collogue and friends for their support, motivation and encouragement during the study.

I would also like to express my gratitude to the examining committee members for their valuable feedback.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
2D	Two Dimensional
3D	Three Dimensional
HMM	Hidden Markov Model
SDK	Software Development Kit
WMV	Windows Media Video
AVI	Audio Video Interleave
BMP	Bitmap

CHAPTER 1

INTRODUCTION

Understanding the human behavior and characteristics is a significant phenomenon in the literature. There have been several studies in the light of human science techniques. In the last decade, the problem has also been addressed with technological improvements. Research area in this field has drawn considerable interest. In this work, we try to detect viewer engagement by automated head motion analysis.

1.1 Purpose and Significance of the Study

Curiosity to know and comprehend much about the environment is one of the most superior goals of mankind. To accomplish this ambition a human being should be aware of not only the others but also the self. This curiosity leads mankind to go for gaining a greater understanding of own behavior. Exploring human mind and understanding human behavior is one of the most essential steps in this way.

Human behavior is the product of a multitude of factors such as attitude, culture, emotions and genetics. People try to sort out the nature of the human behaviors. We ponder about the meaning of life and try to grasp the motivations behind the actions of individuals. As the days go on we figure out that there is a reason lying behind the behavior.

The key element to comprehend human nature profoundly is identifying the main enthusiasms and fascinations of human beings. To state this in different manner, human behavior to some extent reflects peoples' interest. Thus, analyzing human behavior may create the information of human interest. Having knowledge of peoples' interest provides valuable information for different fields. One of the conspicuous fields is education.

The interest of human takes a fundamental place in educational context. One of the most substantial problems in education is the lack of interest which may arise from different subsets of motivation. There are plenty of factors affecting this individual motivation such as discouraging experiences, emotional state, dislike of the content and dislike of teaching method. Some of these factors are related by learners' own state on the other hand, some are under the thumb of instructors. The benefit of this study is making it possible to evaluate the learners' interest by changing the factors in the grip of the instructors. Exploring and practicing some strategies; instructors may increase participants' interests.

Technology plays an increasingly crucial role in numerous fields and among them education benefits one of the biggest. The education system has been integrated with the technology constantly. There are smart boards, e-books, distance education courses and a good deal of applications in this particular field. While some of the applications focus on the content of education; some focus analysis of the instructor or learner itself. One of the most extensive instruments to analyze instructor of learner is video cameras.

There has been a remarkable increase in the extent of areas which are tracked by video cameras. The raw data achieved from cameras can be converted to knowledge by analyzing or interpreting the video content. This conversion process occasionally is beyond the human's own intelligence. Therefore; video analysis techniques are used to acquire knowledge from video raw data, such as abnormal human activities, object recognition, human behavior analysis, face detecting and face tracking.

One of the most considerable subjects of the video analysis is human body parts. They provide a great deal of explanation about human activities. Among the human body parts, face gives more specific details about human intention and emotion. The research areas like face detecting and face tracking comes into prominence in the context of understanding human behavior. Therefore, the interest of the people while in the learning process is directly related to facial features.

The face tracking issue has been studied for a long time by several authors. Most application areas are safety&security and human computer interaction. Face tracking may acquire a wide variety of features such as gestures and facial expressions, eye focusing and head movements. This study concentrates on the head movements feature.

The purpose of this study is to analyze individual head movement while watching educational videos in order to detect viewer engagement level. Viewer engagement corresponds to interest information. If this interest information gathered then the instructor may change their teaching method in order to find a best fit teaching method and design video content that increase student learning outcomes. The result of this study may contribute to the existing literature when investigating learner interest.

This thesis comprised of five chapters. The remaining four chapters are organized as follows:

In Chapter 2, the literature review is presented. Firstly published data sets in literature are presented. Secondly, classification methods are mentioned. Lastly, related works on this subject are detailed.

In Chapter 3, the methodological approaches employed to test research hypotheses are described such as definition of features, feature extraction, N fold cross validation, Hidden Markov Model and evaluation criteria.

In Chapter 4, the experiment is explained. The data set is represented in this chapter. Finally, results of the experiment are discussed.

In Chapter 5, the conclusion of the thesis work is given and possible future work related to the proposed approach is mentioned.

CHAPTER 2

LITERATURE REVIEW

In this chapter the literature review is presented. Firstly, previous studies in this are summarized in related works. In chapter 2.1 Related Work, the aim of the study, employed models, results are indicated. In addition to this, the datasets of the studies are examined by their dataset size, number of frames in video, the video resolution and specific information about the dataset. Secondly, we compare the methods used in these studies.

2.1 Related Work

Gunduz et al. aim to define a method which models temporal behaviors and finds the anomaly behaviors by using stochastic graphical modeling in their study [1]. They employ *Coupled Hidden Markov Model (CHMM)* in order to modeling the temporal behavior and performing global anomaly detection. They use UCSD anomaly detection dataset [2]. They find out that the method which use sparse representation performs faster than the algorithm which uses spatiotemporal features.

Gunduz et al. use UCSD anomaly detection dataset. The dataset comprises of pedestrian walkways scenes which are recorded by a stationary camera mounted at an elevation. It includes both normal and anomaly objects. In a normal setting, video only contains pedestrians while in the anomaly settings video contains bikers, skaters, and small carts beside pedestrians. Each video constitute with 200 frames. The UCSD Ped1 dataset has 34 abnormal training videos while it also has 36 test videos. The resolution of the video is 238x158.

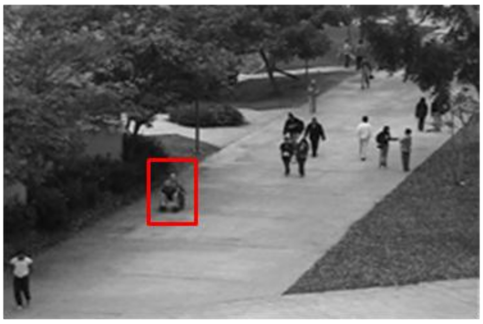


Figure 1: A Snapshot from UCSD Dataset [2]

Sigal et al. propose a novel dataset in order to solve the challenges in gathering ground truth data, especially in 3D videos [3]. The dataset is called HumanEva-I which consist of different objects taking predefined actions such as walking, jogging, gesturing, throw/catch, boxing and combo. The aim is to become a standard dataset for pose estimation and human motion studies.

Dong et al. propose a novel motion representation and apply it to the behavior recognition and abnormality detection [4]. They use both Adaboost and their proposed method which is called *Pointwise motion image (PMI).* They use walking behavior database, action database and KTH database. Walking behavior database and action database is created by themselves. The results show that the success rate of behavior recognition and abnormality detection is high and it is capable of dealing motions with different speeds.

Dong et al. have 3 datasets. Walking Behavior Database contains six kinds of walking in an office scene. Each kind of walking performed by 4 people 25 times. It was recorded by a digital camera and converted into 300x240 BMP file. Action database contains 8 different actions performed by 8 different people 15 times. 960 action sequences are converted into 160x120 BMP file. Lastly, KTH database contains 6 kinds of actions performed many times with 25 people. The actions of KTH database comprise of walking, jogging, running, hand waving, clapping and boxing [5]. All the video have 25fps frame rate and average length is approximately 4 seconds.

Yao et al. aim to detect anomalies in network traffic [6]. They employed a combination of random forest and entropy measure. They used DARPA Intrusion Detection Evaluation Data Set. The result of this study is 99% recall and 48% precision.

Lee et al. propose general purpose of *human motion analysis* method [7]. In the study, in order to learn and classify multi-layer perceptron neural network with a single hidden layer is used. They create their own database. The results show that proposed method has very high accuracy classification.

Lee et al. have 18 video clips which have a resolution of 240x320 pixels and sampling rate of 15fps [7]. Each video stream is approximately 30 seconds length. There are three types of scenarios: walking, running and struggling.

Zhu et al. aim to recognize multiple view human actions [8]. They extracted the silhouette and then apply a novel voting strategy. In the study, random forest classifier is used. The accuracy result is 88%.

Zhu et al. use Inria Xmas Motion Acquisition Sequences (IXMAS) dataset [8]. This dataset is multi-view dataset for human action recognition. Dataset contains 11 actions such as get up, cross arms and pick up performed by three times with 10 different people.

Kakavand et al. aim to detect anomalies by considering network security [9]. They present text mining based anomaly detection model. They employ Mahalanobis Distance Map (MDM) and Support Vector Machine (SVM). The accuracy results of MDM and SVM are 97.44%, 97.45% respectively.

Kakavand et al. use Information Security of Excellence (ISCX) dataset [9]. Dataset contains different kinds of attack such as dos, scan and R2L. Dataset has almost 1512000 packets. Training set is 80% of the set and remaining part is testing set.

Bashir et al. propose a classification system in order to recognize object activity [10]. They segment the trajectories at points of change in curvature and all the sub-trajectories are represented by their principal component analysis. Their first experiment is with *Gaussian Mixture Model* and it failed to capture temporal relations and ordering between subjacent entities. After they use Hidden Markov Model, they have a successful result. They used two datasets: UCI-KDD dataset and Columbia University's Digital Video and Multimedia Group.

Bashir et al. used two different dataset [10]. Australian Sign Language dataset is obtained from UCI-KDD dataset [11]. These trajectories are gathered by registering hand coordinates at each successive instant of time. Second dataset is Columbia University's Digital Video and Multimedia Group and consist of object trajectories tracked from video streams of sport activities such as high jump, slalom skiing, etc [12]. The dataset has approximately 40 high jump trajectory and 68 trajectory of slalom skiing.

Mohamad et al. aim to identify the most appropriate image attributes used by human experts when they select images which contain persons [13]. They define a set of criteria which is able to identify images that contains fine visual of a person. As a result, they obtained three tiers attribute filter.

Brand et al. aim to classify two-handed actions [14]. They use Hidden Markov Model, Linked Hidden Markov Model and Coupled Hidden Markov Model. They use *Viterbi* algorithm to find the maximum likelihood. They compare results and the accuracy rates of Linked HMM, Single HMM and Coupled HMM are 36.5%, 69.2%, and 94.2% respectively.

Brand et al. gathered 3D hand tracking data for three T'ai Chi gestures [14]. They collect 52 sequences. Extracted feature vectors are 3D centroid positions of each of blob positions which characterize hands.

Mehran et al. propose a new method in order to detect the abnormal behaviors and localize it in crowd videos [15]. The path of each person is defined as the total value of social interaction power and target path. They also use particle advection using the average optical flow field. They detect abnormal behaviors without the need of tracking people one by one.

Mehran et al. use UMN dataset and Web dataset [15] . UMN dataset compose of 11 videos [16]. The Web Dataset is composed of available internet sites. There are 20 videos, 12 of them normal, 8 of them abnormal. Video frame width is fixed to 480 pixels.

Ofli et al. propose new representation in order to recognize human actions [17] . They select the minimum skeletal joints which interprets maximum information related with the action by considering mean or variance of joint angle trajectories. They use K-nearest neighbors (KNN) and support vector machine as a classifier.

Ofli et al. use Berkeley multimodal human action database (BMHAD), Motion capture database (HDM05) and MSR Action 3D database. BMHAD contains 11 actions performed by 12 people [17]. There are 659 action sequences totally in this dataset. Some of the actions in this database are jump, bend, throw, sit down and punch. HM05 dataset contains 16 actions performed by 5 different people. There are 393 action sequences in the dataset. Some of the actions are sneaking, sitting down, jogging and kicking forward. MSR Action 3D contains 17 actions performed by 8 different people. There are 379 action sequences.

Celiktutan et al. aim to define people in terms of 5 predefined characters: responsibility, coherence, neuroticism, open-minded and extrovert [18]. They used support vector regression. They have a high success rate for extrovert and neuroticism. They also declare that the most predictable character is extrovert.

Celiktutan et al. use SEMAINE dataset [18]. SEMAINE dataset involves video streams which are interaction between a human and different virtual character [19]. The dataset is composed by 10 person and 3 different half automated characters. Hence, there exist 30 different videos. Each video is clipped to 60 seconds. Video records are evaluated by 21 people, mostly doctorate students.

Fathi et al. represent a method in order to recognize human actions [20]. They create mid-level human features that are extracted from low-level optical flow information. They employed Adaboost as a classifier. The accuracy result of the study is 90.5% for KTH dataset 100% for Weizmann dataset and 71% for Soccer dataset.

Fathi et al. used 4 different datasets which are KTH dataset, Weizmann dataset, soccer dataset and ballet dataset. We have already explained KTH dataset.

Weizmann dataset contains 93 human action videos with 180 x 144 pixels. There are 9 different people and 10 different actions. Thirdly, soccer dataset contains 66 video sequences and 8 different actions. Lastly, ballet dataset has 44 videos and 8 different actions. There are two men and a woman.

Yao et al. propose a fuzzy logic-based system using machine vision in order to recognize human behavior [21]. They employ fuzzy c-means clustering. The result of the study is 94.03% for Weizmann dataset which is also used in the study of Fathi et al. [20].

Wu et al. aim to recognize a set of human activities [22]. They use RGBD camera which is *Kinect* in order to extract motion and body features. They combined SVM model and HMM model. The combined model is employed. The accuracy result of the study is 98.22%.

Wu et al. used daily activity 3D dataset which is published by Cornell University Robot Learning Lab [22]. The dataset has 12 different actions in 5 different scenes which are kitchen, bedroom, office, bathroom and living room. Some activities are brushing teeth, drinking water, talking on the phone and cooking.

Nie et al. aim to recognize human actions [23]. They capture both global and local dynamics of joint trajectories. They employ the combination of Gaussian-Binary restricted Boltzmann machine (GB-RBM) and Hidden Markov Model. The result of this study is 93.1%, 86.4%, 80.2% for MSRC-12 dataset, G3D dataset and MSR Action 3D dataset respectively.

Nie et al. used three different datasets in the study: MSRC-12, G3D and MSR Action 3D [23]. MSRC-12 dataset comprises of 594 sequences from 30 different people. There are 12 gesture types. G3D dataset contains 20 gaming actions captured by *Kinect*. MSR Action 3D has 567 sequences of 20 actions.

Roh et al. propose a dynamic Bayesian network model in order to recognize human gestures and understand sign language [24]. The result of this study is 98% for recognizing human gestures and 94.6% for understanding sign language.

Roh et al. used Korea University Gesture Database (KUGDB) for gesture recognition and American Sign Language Database (ASLDB) of Boston University for sign language understanding [24]. KUGDB has 5 different gesture types. There are 50 samples. Leave one out method is used. ASLDB has video streams for 48 words.

Locke et al. aim to localize the hazardous particle rate in an area [25]. In order to measure the hazardous particle rate they use sensors. They employ SVM as a classifier. The accuracy results depend on the number of sensor and they find high accuracy rate when they increase the number of sensor in environment.

As a result, the all summarized studies above try to analyze picture/video content. Some of these studies focus on analyzing human behavior. The difference between summarized studies and this study is the focus point which is detection automated viewer engagement level.

2.2 Classification of Methods

In this chapter, we declare the employed method of the study and categorize each study by its field of study. We also compared study methods. The information about studies is shown in Table 1: Methods in Surveyed Studies.

Table 1: Methods in Surveyed Studies

Study Reference	Method Name	Field of Study
Gunduz et al. [1]	Coupled HMM	Anomaly Detection
Dong et al. [4]	Adaboost and PMI	Action Recognition, Anomaly Detection
Yao et al. [6]	Random Forest	Anomaly Detection
Lee et al. [7]	A Novel Method	Anomaly Detection
Zhu et al. [8]	Random Forest	Action Recognition
Kakavand et al. [9]	SVM and MDM	Anomaly Detection
Bashir et al. [10]	HMM	Sign Language Understanding
Mohamad et al. [13]	Data Triangulation Method	Identifying Human Attributes
Brand et al. [14]	HMM&LinkedHMM&Coupled HMM	Action Recognition
Mehran et al. [15]	Social Force Model	Anomaly Detection
Ofli et al. [17]	SVM and KNN	Action Recognition
Celiktutan et al. [18]	Support Vector Regression	Character Analyze
Fathi et al. [20]	Adaboost	Action Recognition
Yao et al. [21]	Fuzzy c-means clustering	Action Recognition
Wu et al. [22]	SVM&HMM	Action Recognition
Ni et al. [23]	GB-RBM&HMM	Action Recognition
Roh et al. [24]	Simplified Dynamic Bayesian Network	Action Recognition Sign Language Understanding
Locke et al. [25]	SVM	Anomaly Detection

As it is shown in Table 1; most popular methods through surveyed studies are Hidden Markov Model, SVM and Random Forest. As, the objective of this research is to detect viewer interest for video basis, not for the frame basis; HMM is not selected as a classification model because each video frame is not analyzed individually and there is no state transition assumption between consecutive frames.

CHAPTER 3

RESEARCH METHODOLOGY

In this thesis, the main aim is to automatically detect the engagement level (engaged or not engaged) of viewers from their headshot videos recorded while they are watching educational videos. Firstly, we use a face detection algorithm which detects faces and record face positions for each frame into a file. Secondly, the head motion features as specified in the planning phase are extracted by using the detected face positions in consecutive frames. Finally, *SVM* and *Random Forest* are used to detect user engagement.

In the experiments, we test various combinations of features to observe their effects on the performance. We use a 10-fold cross validation scheme where the model is trained using nine out of ten subsamples of the dataset. The trained model is tested on the remaining one out of ten subsamples of the dataset. Training and testing phase are performed 5 times for each selected feature variables and the average of these 5 iterations is calculated. Evaluation criteria are calculated by the average iteration result of interest. Stop criteria of this process are determined by considering accuracy distributions. After obtaining the entire interest test result for given feature; all interest result is compared and the best result is documented. This process is explained in Figure 3 and it is detailed in the following sections.

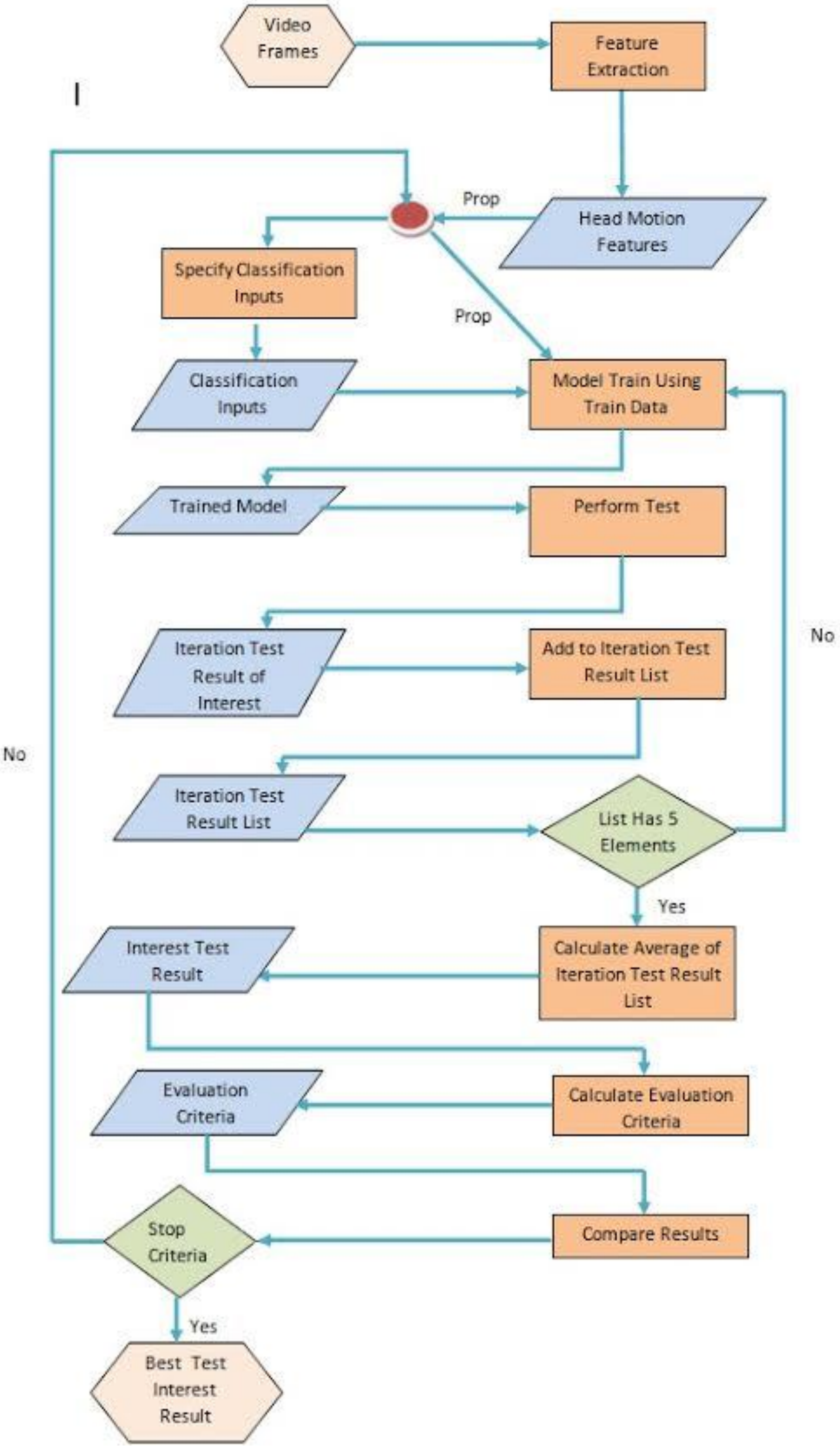

Video Frames
Feature Extraction
Prop
Head Motion Features
Specify Classification Inputs
Prop
Classification Inputs
Model Train Using Train Data
Trained Model
Perform Test
Iteration Test Result of Interest
Add to Iteration Test Result List
No
No
Iteration Test Result List
List Has 5 Elements
Yes
Interest Test Result
Calculate Average of Iteration Test Result List
Evaluation Criteria
Calculate Evaluation Criteria
Stop Criteria
Compare Results
Yes
Best Test Interest Result

Figure 2: Process flow of the proposed system

3.1 Feature Extraction

We detect the location of the face at each frame. These localizations can be used to characterize the head movements. The localization of the head contains the center position of the head in 2 dimensions and the radius of the head.

Interpreting the positions of x-y coordinates and radius of the head in the most correct way was one of the key points in this study. In each subsequent frame, these positions were compared. For each successive frame, changes were gathered. This information was written into a file. To sum up, head position change in the x dimension, head position change in y dimension and the change of radius of the head is calculated. Overview of the code is as follows:

- Read from file
- Extract x, y positions and radius of faces for each frame
- Calculate changes of x, y and radius of face for each frame

Instead of specifying classification input data as head position changes of every frame for every person, we specified input parameters as mean value of absolute changes of head in the x dimension, mean value of absolute changes of head in the y dimension, mean value of the head radius and the radius of the head. The calculations of these values are shown in 4.2.3.

3.2 Classification Models

Random forest is a classification model which constructs multiple decision trees at training time. Each decision tree is a different subset of training set. When testing the data; result is determined by considering the result of each decision tree for that specific data.

Support Vector Machine (SVM) is a non-probabilistic binary classification model. Due to we have two class in this study; it is a proper classification model. Sequential Minimal Optimization (SMO) is an efficient algorithm for SVM. It is fast training of SVM. SMO is widely used for training SVM so that we used SMO for training.

In training and testing phases, we used a tool named *Weka* data mining software tool [26]. Weka Data Mining Software Tool supports a set of data mining tasks such as preprocessing, feature selection, classifying, clustering and visualization. Weka contains various types of classification models such as Random Forest, Support Vector Machines and NaiveBayes.

We train the SVM and Random Forest models separately to test the performance. We classified each of the testing data to following states: “person watching the video with interest” or “person watching the video without interest”.

3.3 N - Fold Cross Validation

Cross validation also known as rotation estimation is a model validation technique. This technique provides evaluating the statistical result of model with independent dataset. Since we do not have unlimited number of examples in our dataset, calculating the error rate with limited dataset does not correspond to true error rate. However, we should train and test in order to obtain error rate. Training and testing using all our datasets and calculating the error rate would not be a good indicator of performance on future data because of overfitting. Splitting the data into train and test could also be misleading as it is dependent on the way the data is split. In order to overcome this problem we used 3-fold cross validation. For each three experiment we use two folds for training and the remaining one for testing. The outline of this procedure is explained as follows:

1. Two out of three available data specified as training set and remaining part as test set
2. Model is trained with training set.
3. The model is tested using test set and error rate is calculated.
4. Steps 2-4 are repeated three times by changing test set as specified
5. Error rates are averaged.

3.4 Evaluation Criteria

In order to visualize the performance of the experiment, confusion matrix is formed. *True Positive* (TP) means actual classification of the instance is interested and predicted classification of instance is also interested. *False Negative* (FN) means actual classification of the instance is interested but predicted classification of the instance is not interested. *False Positive* (FP) means actual classification of the instance is not interested while the predicted classification of the instance is interested. *True Negative* (TN) corresponds to actual and predicted classifications of instances are not interested. Confusion matrix is as follows:

Table 2: Confusion Matrix

		Predicted Classes	
		Interested	Not Interested
Actual Classes	Interested	TP	FN
	Not Interested	FP	TN

Confusion matrix is generated for each classification model. In addition to this; the experiment is repeated by restricting the properties of head movements. Confusion matrix is calculated for every experiment as well. An example of the confusion matrix is illustrated in Table 3.

Table 3: Confusion Matrix of 3D (Mean+Variance) for Random Forest

		Predicted Class	
		Interested	Not Interested
Actual Classes	Interested	145	20
	Not Interested	25	140

One of the significant evaluation criteria is *accuracy*. Accuracy is the proportion of the true result.

$$P = TP + FN \tag{3.1}$$

$$F = FP + TN \tag{3.2}$$

$$Accuracy = \frac{TP + TN}{P + N} \tag{3.3}$$

Another evaluation criterion is *precision* which also known as positive predictive value.

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

Recall also known as *sensitivity* is *True Positive* rate. This criterion measures the proportion of actual positives which are correctly identified.

$$Recall = \frac{TP}{TP + FN} \tag{3.3}$$

Specificity is *True Negative* rate. This criterion measures the proportion of actual negative which are correctly identified.

$$Specificity = \frac{TN}{FP + TN} \tag{3.3}$$

Fall-out is *False Positive* rate.

$$Fall - out = \frac{FP}{FP + TN} \tag{3.3}$$

Miss-rate is *False Negative* rate.

$$Miss - rate = \frac{FN}{FN + TP} \tag{3.3}$$

3.5 Limitations

Limitations of this study are minimum distance, maximum distance and number of person in one video.

- Minimum distance between camera and participant should be 30cm.
- Maximum distance between camera and participant should be 120cm.
- Only one person should be in the video.

CHAPTER 4

EXPERIMENTS& RESULTS

In this chapter, we explain the datasets, experiments and their results.

4.1 Data Set

For this research, we recorded a dataset for experimentation. Our dataset is composed of headshot videos of people recorded while they were watching educational videos. Educational videos were chosen from the program *Wissen Vor Acht*. Yet German is the original language of the program and the mother tongue of all the participants in the research is Turkish; the Turkish translated version of the program *İki Dakikada Bilim* was chosen.

İki Dakikada Bilim started in 2008 has hundreds of educational videos in a variety of fields such as sociology, mathematics, biology, economics and earth sciences. The average duration of videos is two minutes. The videos were selected by the field of interest of the participants. The participants were requested to select two videos; a video that they find interesting and would like to watch and a video they do not find interesting. The videos were listed to the people through the internet site [27]. People choose the video by looking at the title and a snapshot of the video. 33 participants joined this research voluntarily. Because each participant watched two videos; there are a total of 66 videos in the dataset.

Gender characteristic, age distribution and education level of participants are shown in the Tables 1, 2 and 3 respectively. The age range of the participants is from 20 to 34 years. The average age of the participants is 26.48. Considering the gender characteristics of the participants, 33% of the participants are female while 67% of the participants are male (11 female and 22 male).Education levels of participants are categorized as shown Table 6. 55% of the participants have a Bachelor’s degree. 33% master degree and 6% are high school graduates. The smallest portion of the participants has a PhD (3%).While 64% of the participants are still student; 36% of them not related with any kind of school or course.

Table 4: Gender Information of Dataset

Gender	Number Of Participants	Percentage
Male	22	33%
Female	11	67%
Total	33	100%

Table 5: Age Information of Dataset

Age Range	Number Of Participants	Percentage
20 -24	7	21.2%
25 – 29	20	60.6%
30-34	6	18.2%

Table 6: Education Level Information of Dataset

Education Level	Number Of Participants	Percentage
Bachelor’s Degree	18	55%
Master’s Degree	12	36%
High School Graduate	2	6%
PhD	1	3%

Table 7: Student Information of Dataset

Current Student Information	Number Of Participants	Percentage
Student	21	64%
Non Student	12	36%

All the videos were recorded by the same video camera and stored in WMV format. There is no corruption or inaccurate records thus there is no need for a data cleansing process.

4.2 Experiment

The experiment comprised of the following steps. The first step in the experiment is the planning step. In this step; we specified the restrictions as well as the hardware components by considering contemporary face detection applications. In this step we also selected the face detection application by taking into consideration of the decision of the first step. The second step is preprocessing. In this step data set is created which is briefly explained in 4.1. In this step data is converted to the appropriate format for the chosen application. The last step is processing the data. In this step we extract that features and train using SVM and Random Forest. Then we tested on the remaining data.

4.2.1 Planning

To detect engagement of persons watching educational videos by automated analysis of head motion, we need to obtain videos with their headshots while watching. Such videos can be obtained by RGB cameras and RGB-D cameras.

RGB cameras are the most ubiquitous camera types and they are available almost everywhere. While their cost range is extensive, cheaper cameras would be sufficient for tracking head movements. The RGB camera does not contain depth information yet three-dimensional information can be obtained by use of multiple cameras. Even though the depth information would have extra benefits, RGB cameras are the most convenient with regard to cost and extensity.

RGB-D camera is the second alternative camera type. This camera type collects the depth information as distinct from RGB camera. There were two main choices to select RGB-D camera: *Kinect* and *Intel® RealSense™ 3D Camera.*

One of the most well-known RGB-D cameras is *Kinect. Kinect* is a motion sensing input device. Despite it was originally designed for playing games, it is used in data acquisition for various research studies as well. Microsoft released *Kinect* SDK (software development kit) that allows developers to write applications. Software technology enables gesture recognition, face tracking and voice recognition. *Kinect* is capable of simultaneously tracking up to six people. It is also possible to track skeleton. However *Kinect* is more expensive compared to RGB cameras and not as commonly available.

Intel® RealSense™ 3D Camera is the second alternative RGB-D camera in order to obtain head movement features. When it comes to face tracking, *Intel® RealSense™ 3D Camera* has a similar performance to *Kinect*. It allows removing the background and claims that it contains advanced gesture recognition technology. Despite all its advantages, *Intel® RealSense™ 3D Camera* is also more expensive compared to RGB cameras and not as commonly available.

As a result, we decided to implement the system using a standard RGB camera so that it can be used without any requirement for special hardware and could easily be installed for multiple users.

Table 8: Comparisons of Hardware Systems

Hardware System	Cost	Extensity	Extras
Kinect	Higher	Limited	SDK, tracks up to six people, track skeleton, gesture recognition technology
Senz3d	Higher	Limited	Removing background, gesture recognition technology, depth information
Digital Camera	Lower	Common	Available face detection applications

The second thing to decide is the number of persons to be detected in one camera shot. There were two possible options: one camera for a group environment such as the classroom or meeting room and one camera for one person.

Detecting engagement of multiple persons using a single camera is less costly. Furthermore; as it is shown in Figure 3, some schools are reported to have plans to place cameras [28]. However, tracking of multiple persons with one camera is more difficult and provide less accurate tracking information and margin of error of such as system would be higher.

Figure 3: One Camera for Multiple Person [28]

Detecting engagement of individuals with one camera for one person has less margin of error compared to one camera for multiple persons. Since the camera is focused on one participant, it collects more accurate data. On the other hand, one camera for one person has a higher cost. However, nowadays every student has their laptops and tablets in the classroom as it is shown in Figure 4. In addition to this, some clubs have tablets attached to the back of each seat and they started using facial analysis technology [29]. Because it is more accurate and cameras are widely available, one camera for one person is selected for this research.

Figure 4: One Camera For One Person [30]

Table 9: Comparisons of Focus Points

Focus Point	Margin Of Error	Cost	Extensity
Multiple Persons	Higher	Lower	Normal
One Person	Lower	Higher	Higher

The face detection method provides the fundamental data for analysis in this work. Hence, selection of a more accurate face detection algorithm is expected to have a positive impact on the performance. The article “*A Benchmark for Face Detection in Unconstrained Settings*” [31] provides a benchmark summary for a number of face detection algorithms. This work was then extended with a number of other algorithms [32]. In the article while different face detection applications were compared in discrete time and in continuous time, since our study is based on the video stream of participants; continuous time comparisons were of interest. The best result in the article belongs to Viola-Jones [33].

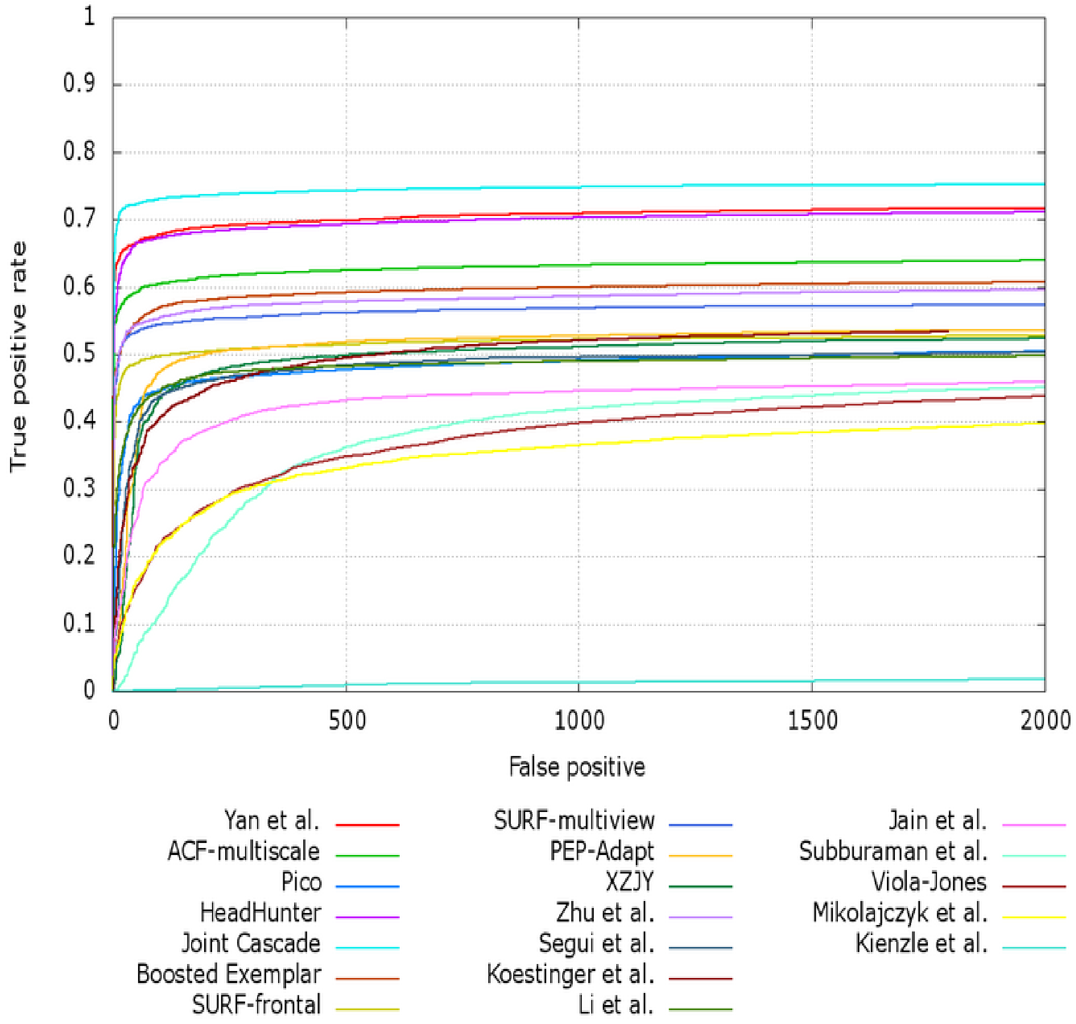


Figure 5: ROC Curves of Face Detection Algorithms Based On Continues Score [32]

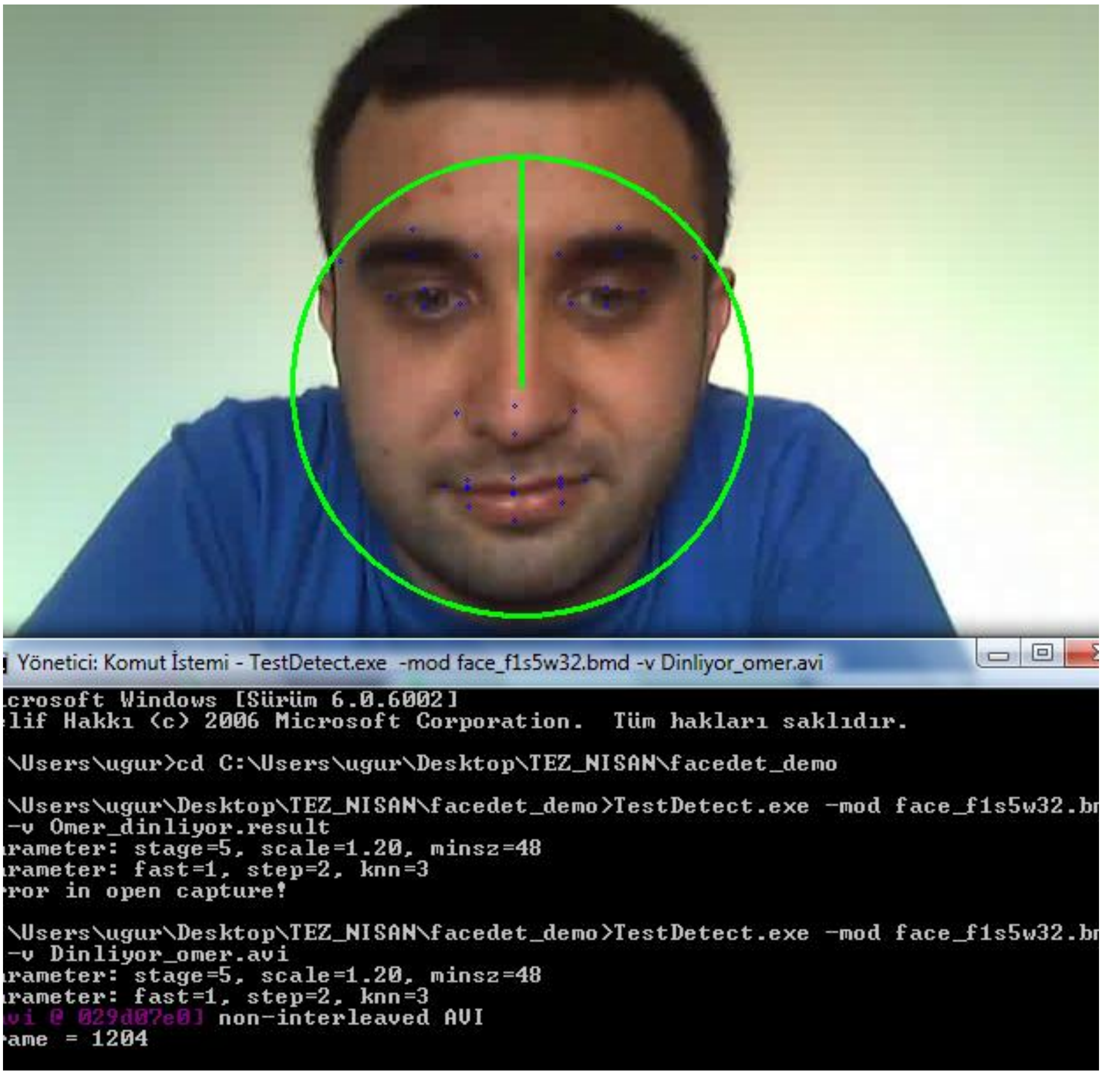
However, the researchers have continuously updated the benchmark result on the web page by adding more recent algorithms as they are available. By the time we developed our system; Li et al. had the best continuous ROC curve [32] and we adopted this algorithm. For future work, more recent face detection algorithm might be integrated in the system.

4.2.2 Preprocessing

In this step, data is created and prepared for the processing step. Data type conversion and transformation into the suitable format are performed.

First of all, we run the face detection application and record its output giving the detected face positions in each frame. Then, we calculated the temporal change of face positions for each frame.

All the video data that contains the participants' behavior is composed by taking into consideration of the face detection application and its restrictions. Chosen face detection application can track faces successfully up to the distance of 1.2 meters. The success rate of tracking the face decreases further this point. Thus, participants were filmed with the distance of approximately 50 cm. While the data was created in various illumination environments the algorithm worked satisfactorily in different conditions. Utilization permit of Figure 6 is at Appendix C: Utilization Permit.



Yönetici: Komut İstemi - TestDetect.exe -mod face_f1s5w32.bmd -v Dinliyor_omer.avi

```
crosoft Windows [Sürüm 6.0.6002]
lif Hakkı (c) 2006 Microsoft Corporation.  Tüm hakları saklıdır.

\Users\ugur>cd C:\Users\ugur\Desktop\TEZ_NISAN\facedet_demo

\Users\ugur\Desktop\TEZ_NISAN\facedet_demo>TestDetect.exe -mod face_f1s5w32.bm
-v Omer_dinliyor.result
rameter: stage=5, scale=1.20, minsz=48
rameter: fast=1, step=2, knn=3
ror in open capture!

\Users\ugur\Desktop\TEZ_NISAN\facedet_demo>TestDetect.exe -mod face_f1s5w32.bm
-v Dinliyor_omer.avi
rameter: stage=5, scale=1.20, minsz=48
rameter: fast=1, step=2, knn=3
vi @ 029d07e0] non-interleaved AVI
ame = 1204
```

Figure 6: A Screenshot from One of the Videos

All the video data was stored in *WMV* format. However, the face detection application expects AVI files as input. Video data was converted into AVI format [34].

Secondly, before composing the video data, we decided the settings of the tracking applications. Settings were as follows:

- [-mod modname] binary-model-name
- [-m] minimal scan window width
- [-fast] fast-level (can be 0, 1, 2)
- [-z] scale in scan-win-size

- ➢ As binary model, we used face_f1s5w32.bmd binary model.
- ➢ Minimal scan window is selected as 32.
- ➢ Fast level is selected as 1.
- ➢ Scale in scan windows size is selected 1.2.

4.2.3 Processing

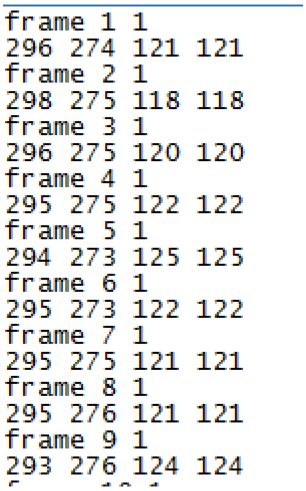
This step consists of features extraction, training and testing.

The outcome of the face detection application is the result file which contains face positions of the participant for each frame. If the application could not find a face in a frame, then it does not log anything to the result file. There were 217450 frames in the dataset. The algorithm was not able to detect the face for 44754 of them (20%). The videos where people are interested are formed of 107062 frames. The algorithm was not able to detect a face in 11860 frames (11%). The videos where people are uninterested are comprised of 110388 frames. The algorithm was not able to detect a face in 32894 of them (30%).

Table 10: Number of Identified/Unidentified Faces in Frames

Dataset	Number Of Frames Face Was Not Detected	Number Of Frames Face Was Detected
Interested	11860	107062
Non-Interested	32894	110388

The outcome of the face detection application contains frame number, number of founded face, x position of the face, y position of the face, x radius of the face and y radius of the face respectively. Figure 7 is the example of the face detection application result file:

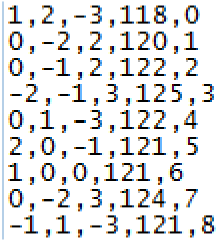


```
frame 1 1
296 274 121 121
frame 2 1
298 275 118 118
frame 3 1
296 275 120 120
frame 4 1
295 275 122 122
frame 5 1
294 273 125 125
frame 6 1
295 273 122 122
frame 7 1
295 275 121 121
frame 8 1
295 276 121 121
frame 9 1
293 276 124 124
```

Figure 7: Face Detection Application Result File

In Figure 7, first line shows frame number and number of faces that found in that frame respectively. Second line indicates x position of the face center, y position of the face center and two times radius of the face respectively. These two lines repeat for each frame.

In order to extract movements of head in each dimension; we developed an application which extracts change of the face position in x dimension x_{change}, change of the face position in y dimension y_{change}, change of the radius of the face z_{change}, the radius of the face r_{head} and the sequence number *seq*. The change of radius is related to displacement in z dimension (z_{change}) because if a participant moves away from the camera, the observed face will be larger in the image (larger radius) and moves closer to the camera, the observed face will be smaller (smaller radius). Figure 8 shows an example part of the result file after executing the application.



```
1,2,-3,118,0
0,-2,2,120,1
0,-1,2,122,2
-2,-1,3,125,3
0,1,-3,122,4
2,0,-1,121,5
1,0,0,121,6
0,-2,3,124,7
-1,1,-3,121,8
```

Figure 8: Movement of Head in Each Frame Result File

Figure 8: Movement of Head in Each Frame Result File columns correspond to x_{change}, y_{change}, z_{change}, r_{head} and frame number respectively.

As the objective of this research is detection of participants' interest not for the frame basis, but for the video basis; the mean value and variance is calculated. However, the mean of values might mislead us because we search for the potential changes of the head without considering any direction. Therefore, the absolute value of each dimension was calculated. The absolute value of variance is calculated as well.

Each participant was filmed with a different distance to the camera. In addition to this, each participant has different head size. So, the minimum change of the participant having larger head size is expected to be different than that of a participant having a smaller head size. This condition also works for the distance. In order to compensate the conditions, the mean value and the variance are normalized by dividing into the square of the mean value of the radius of the face.

While r_{head} is the face radius; mean value of the radius of the face is calculated as follows where f is the number of frames in the video:

$$\mu_{hsize} = \frac{\sum_f r_{head}}{f} \tag{4.1}$$

After obtaining the mean value of the radius of face; the mean absolute values of each dimension are calculated.

Mean absolute value of x dimension is calculated as follows:

$$\mu_{absx} = \frac{\sum_f |x_{change}|}{{\mu_{hsize}}^2 * f} \tag{4.2}$$

Mean absolute value of y dimension is calculated as follows:

$$\mu_{absy} = \frac{\sum_f |y_{change}|}{{\mu_{hsize}}^2 * f} \tag{4.3}$$

Mean absolute value of z dimension is calculated as follows:

$$\mu_{absz} = \frac{\sum_f |z_{change}|}{{\mu_{hsize}}^2 * f} \tag{4.4}$$

Then, variance of absolute values of x dimension is calculated as follows:

$$\sigma^2{}_{absx} = \frac{\sum_f (|x_{change}| - \mu_{absx})^2}{{\mu_{hsize}}^2 * f} \tag{4.5}$$

Variance of absolute values of y dimension is calculated as follows:

$$\sigma^2{}_{absy} = \frac{\sum_f (|y_{change}| - \mu_{absy})^2}{{\mu_{hsize}}^2 * f} \tag{4.6}$$

Variance of absolute values of z dimension is calculated as follows:

$$\sigma^2{}_{absz} = \frac{\sum_f (|z_{change}| - \mu_{absz})^2}{{\mu_{hsize}}^2 * f} \tag{4.7}$$

After these calculations, mean values and variances of each 3 dimension was gathered. These numbers summarize the movement of the head throughout the video stream.

Input variables of classification models are calculated mean values and variances. μ_{absx}, μ_{absy}, μ_{absz}, $\sigma^2{}_{absx}$, $\sigma^2{}_{absy}$ and $\sigma^2{}_{absz}$ variables were calculated separately for each 33 participants. In order to give an indication of how well the model learn when it is asked to identify the interest for the data it has not already seen, testing data was not used while the model was training. While N Fold Cross Validation is used, the original sample is randomly partitioned into three equal size subsamples. Each time

44 videos are used in training and 22 videos is used in testing. 3-fold cross validation is performed as follows:

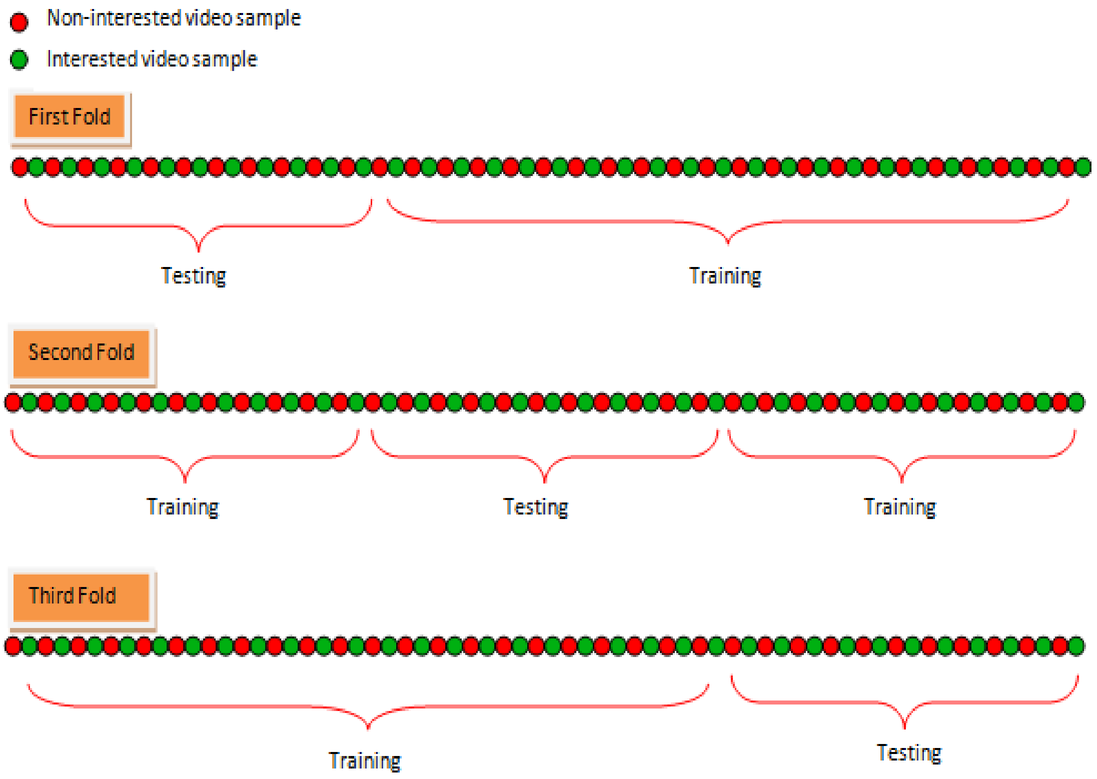


Figure 9: 3 Fold Cross Validation

As the Random Forest classification will be highly random; we tried to stabilize the outcome. By multiple observations of *Random Forest* training and testing phases, we maximized the deterministic of the model. We observed 5 times for each training and testing.

We also experiment using different combinations of the extracted features. The first experiment was formed of mean values and variance of three dimensions. In the pursuit of this experiment, only mean values of three dimensions and only variances of three dimensions were observed separately. In addition to this, all the mentioned experiments were observed in two dimensions by eliminating the z dimension.

4.3 Results & Discussions

In this part, the results that are obtained from testing phase are presented. Evaluation metrics that are used to appraise the proposed system are demonstrated.

All the metric results given in this topic; are visualized in Appendix B: Visualization of Evaluation Criteria Results.

Firstly, we measure the accuracy metric. Due to having a balanced dataset, accuracy metric which indicates the number of correct predictions to the total predictions, is the most valuable metric. All the accuracy distributions are located in Table 11.

Table 11: Accuracy Results

	2D (Mean)	2D (Variance)	2D(Mean+ Variance)	3D (Mean)	3D (Variance)	3D(Mean+ Variance)
Random Forest	0.833	0.848	0.833	0.864	**0.894**	0.864
SVM	**0.879**	0.667	0.848	0.864	0.773	0.833

As it is shown in Table 11, the best accuracy result is obtained by 3D (Variance) of Random Forest classification. This corresponds to; mean values of head movement mislead us when classifying engagement of the viewer. In addition to this, third dimension feature of head movement in the video stream is a fundamental feature. Also, Random Forest classification model has better results than SVM. SVM has a better result only for the 2D mean value feature and 2D (Mean+Variance) feature. For 3D Mean features SVM and Random Forest has same results.

The best accuracy result for SVM is 2D Mean values which is 0.879. Worst accuracy result belongs to SVM which is 0.667. Worst accuracy result of Random Forest has better than mean value of SVM results 0.833> 0.810.

Recall metric also known as Hit rate or sensitivity is another significant metric. It corresponds to true positive rate. The recall results are shown in Table 12.

Table 12: Recall Results

	2D (Mean)	2D (Variance)	2D(Mean+ Variance)	3D (Mean)	3D (Variance)	3D(Mean+ Variance)
Random Forest	0.848	0.848	0.818	0.879	**0.909**	0.879
SVM	0.909	**1.000**	0.909	0.909	0.970	0.879

As shown in Table 12, the best result belong to 2D (Variance) of SVM which is 1.Although this is the best recall result, 2D (Variance) of SVM classification has lower accuracy result (0.667).The best recall result of Random forest is 0.909 which belongs to 3D variance feature. The best accuracy result which is 3D (Variance) with Random Forest also has the best recall value of Random Forest. Results also show that, Variance values have better performance over Mean values when calculating recall rate.

The worst recall result which is 0.818 belongs to Random Forest for 2D (Mean+Variance). Worst recall result of SVM is 0.879 for 3D (Mean+Variance) feature.

Specificity which is true negative rate, usually calculated with recall in order to measure prediction of the model. Specificity results of the features for given parameters are shown in Table 13.

Table 13: Specificity Results

	2D (Mean)	2D (Variance)	2D(Mean+ Variance)	3D (Mean)	3D (Variance)	3D(Mean+ Variance)
Random Forest	0.818	0.848	0.848	0.848	**0.879**	0.848
SVM	**0.848**	0.333	0.788	0.818	0.576	0.788

As shown in the Table 13, the best specificity result belongs to 3D (Variance) for Random Forest. The best accuracy result which is 3D (Variance) of Random Forest has the best specificity result.

The specificity interval for Random forest has lower than the interval for SVM. In addition to this, worst specificity results which belongs to SVM 2D Variance also has the best recall result.

Precision which is the true negative rate, usually calculated with recall in order to measure prediction of the model. Specificity results of the features for given parameters are shown in Table 14.

Table 14: Precision Results

	2D (Mean)	2D (Variance)	2D(Mean+ Variance)	3D (Mean)	3D (Variance)	3D(Mean+ Variance)
Random Forest	0.824	0.848	0.844	0.853	**0.882**	0.853
SVM	**0.857**	0.600	0.811	0.833	0.696	0.806

As shown in Table 14, the best precision result belongs to 3D (Variance) for Random Forest. The best accuracy result which is 3D (Variance) has also the best precision value. In addition to this, worst precision value belongs to 2D (Mean) of SVM which is 0.600.

Fall-out is also known as false positive rate, closely related with specificity. Sum of each specificity and fall-out values equals to 1. Table 15 shows the result on the basis of features and given parameters.

Table 15:Fall-out Results

	2D (Mean)	2D (Variance)	2D(Mean+ Variance)	3D (Mean)	3D (Variance)	3D(Mean+ Variance)
Random Forest	0.182	0.152	0.152	0.152	**0.121**	0.152
SVM	**0.152**	0.667	0.212	0.182	0.424	0.212

As shown in the Table 15, the best result is the lowest value 0.121 which is 3D (Variance) for Random Forest. The best SVM fall-out result is 0.152 which belongs to 2D Mean feature.

Worst fall out result is 0.667 with SVM classification for 2D Variance.

Miss-rate is also known as false negative rate, closely related with recall. Sum of each recall and miss-rate values equals to 1. Table 16 shows the miss-rate results on the basis of features and given parameters.

Table 16: Miss-rate Results

	2D (Mean)	2D (Variance)	2D(Mean+ Variance)	3D (Mean)	3D (Variance)	3D(Mean+ Variance)
Random Forest	0.152	0.152	0.182	0.121	**0.091**	0.121
SVM	0.091	**0.000**	0.091	0.091	0.030	0.121

As shown in the Table 16, the best result is the lowest value 0 which belongs to 2D (Variance) of SVM classification.

The best accuracy result which is 3D (Variance) of Random Forest has 0.091 miss-rates which is also the best miss-rate result of Random Forest.

CHAPTER 5

CONCLUSIONS AND FUTURE WORKS

In this last chapter of the thesis, the results of our experiments and contributions are concluded. We also discuss the limitations of our study and present the future study directions.

5. 1 Conclusion

In this study, we detected viewer engagement level by using Random Forest and SVM classification models with head motion analysis from video stream in *Weka* environment. To achieve this goal, the private dataset is comprised of 66 instances in this study as well. According to the classification results which are produced by classification models, the viewer engagement levels are calculated in terms of accuracy of 89.4% shown in Figure 11: Accuracy Results of Features. This result shows that viewer engagement levels can be detected with a high reliable rate. Due to the measurable viewer engagement level, the quality level of the educational videos might increase by using these findings. In addition to this, the viewer can acquire videos that he/she is not engaged.

5.2 Further Research

In future works, nominal scaling for viewer engagement level can be detailed such as not engaged, less engaged, unclear, engaged, very engaged. This study extracts the viewer engagement of only one person at the same time. The model which is employed in the study can be refined by analyzing multiple viewers at the same time. Furthermore, the future research might expand the dataset.

REFERENCES

[1] A. E. Gunduz, T. T. Temizel and A. Temizel, "Pedestrian zone anomaly detection by non-parametric temporal modelling," in *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, 2014.

[2] February 2013. [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm. [Accessed 2 January 2015].

[3] L. Sigal and M. J. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," *Brown Univertsity TR,* vol. 120, 2006.

[4] Q. Dong, Y. Wu and Z. Hu, "Pointwise motion image (PMI): A novel motion representation and its applications to abnormality detection and behavior recognition," *Circuits and Systems for Video Technology, IEEE Transactions on,* vol. 19, pp. 407--416, 2009.

[5] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004.

[6] D. Yao, M. Yin, J. Luo and S. Zhang, "Network Anomaly Detection Using Random Forests and Entropy of Traffic Features," in *Multimedia Information Networking and Security (MINES), 2012 Fourth International Conference on*, 2012.

[7] C.P. Lee, K. M. Lim and W.L. Woon, "Statistical and entropy based multi purpose human motion analysis," in *Signal Processing Systems (ICSPS), 2010 2nd International Conference on*, 2010.

[8] F. Zhu, L. Shao and M. Lin, "Multi-view action recognition using local similarity random forests and sensor fusion," *Pattern recognition letters*, vol. 34, pp. 20-24, 2013.

[9] M. Kakavand, N. Mustapha and A. Mustapha, "A Text Mining-Based Anomaly Detection Model in Network Security," *Global Journal of Computer Science and Technology,* vol. 14, 2015.

[10] F. I. Bashir, A. A. Khokhar and D. Schonfeld, "Object Trajectory-Based Activity Classification and recognition using hidden Markov models," *Image Processing, IEEE Transactions on,* vol. 16, pp. 1912--1919, 2007.

[11] S. Hettich and S. D. Bay, 1999. [Online]. Available: http://kdd.ics.uci.edu/. [Accessed 2 January 2015].

[12] W. Chen and S. F. Chang, "Motion trajectory matching of video objects," in *International Society for Optics and Photonics*, 1999.

[13] N. Mohamad, M. Annamalai and S. S. Salleh, "A knowledge-based approach in video frame processing using Iterative Qualitative Data Analysis," in *Semantic Technology and Information Retrieval (STAIR), 2011 International Conference on*, 2011.

[14] M. Brand, N. Oliver and A. Pentland, "Coupled hidden Markov models for complex action recognition," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, 1997.

[15] R. Mehran, A. Oyama and M. Shah, "Abnormal crowd behavior detection using social force model," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on,* 2009.

[16] [Online]. Available: http://mha.cs.umn.edu/movies/crowdactivity-all.avi. [Accessed 2 January 2015].

[17] F. Ofli, R. Chaudhry. G. Kurillo and R. Vidal, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation,* vol. 25, pp. 24--38, 2014.

[18] O. Celiktutan and H. Gunes, "Continuous prediction of trait impressions," in *Signal Processing and Communications Applications Conference (SIU), 2014 22nd,* 2014.

[19] G. McKeown, M. Valstar, R. Cowie, M. Pantic and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between

a person and a limited agent," in *Affective Computing, IEEE Transactions on*, 2012.

[20] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.

[21] B. Yao, H. Hagras, M. J. Alhaddad and D. Alghazzawi, "A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments," *Soft Computing,* pp. 1--8, 2014.

[22] H. Wu, W. Pan, X. Xiong and S. Xu, "Human activity recognition based on the combined SVM&HMM," in *Information and Automation (ICIA), 2014 IEEE International Conference on*, 2014.

[23] S. Nie and Q. Ji, "Capturing Global and Local Dynamics for Human Action Recognition," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*, 2014.

[24] M. C. Roh and S. W. Lee, "Human gesture recognition using a simplified dynamic Bayesian network," *Multimedia Systems,* pp. 1--12, 2014.

[25] R. T. Locke and I. C. Paschalidis, "Detection and localization of harmful atmospheric releases via support vector machines," *Environmental Systems Research,* vol. 4, pp. 1--11, 2015.

[26] Machine Learning Group at the University of Waikato, "Weka 3: Data Mining Software in Java," Univercity of Waikato, [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed 1 April 2015].

[27] [Online]. Available: http://alkislarlayasiyorum.com/etiket/?kelime=iki+dakikada+bilim. [Accessed 2 January 2015].

[28] D. R. Holmes, May 2012. [Online]. Available: http://theholmeseducationpost.com/2012/05/is-it-time-to-place-cameras-in-the-classroom/. [Accessed 2 January 2015].

[29] J. Wakefield, October 2014. [Online]. Available: http://www.bbc.com/news/technology-29551380. [Accessed 1 December 2014].

[30] [Online]. Available: http://1to1.eun.org/web/acer. [Accessed 2 January 2015].

[31] V. Jain and E. G. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," *UMass Amherst Technical Report*, p. 8, 2010.

[32] A. R. Chowdhury, May 2010. [Online]. Available: http://vis-www.cs.umass.edu/fddb/results.html. [Accessed 3 January 2015].

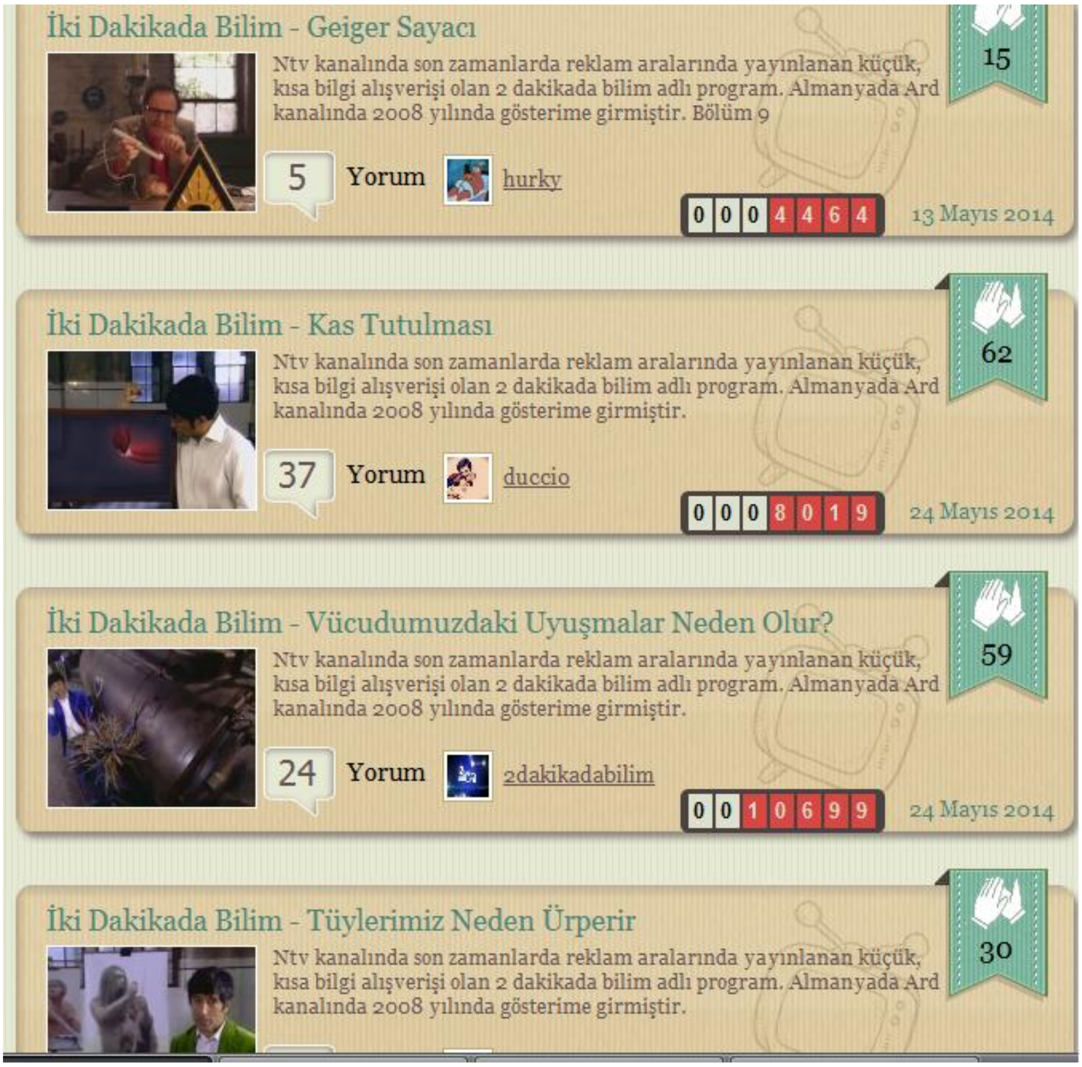
[33] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, pp. 137--154, 2004.

[34] [Online]. Available: http://video.online-convert.com/convert-to-avi. [Accessed 12 November 2015].

[35] [Online]. Available: https://www.reallusion.com/iclone/help/iclone5/PRO/08_Animation/Device_Mocap/Capturing_Head_Rotation.htm. [Accessed 2 January 2015].

APPENDICES

Appendix A:



İki Dakikada Bilim - Geiger Sayacı

Ntv kanalında son zamanlarda reklam aralarında yayınlanan küçük, kısa bilgi alışverişi olan 2 dakikada bilim adlı program. Almanyada Ard kanalında 2008 yılında gösterime girmiştir. Bölüm 9

15

5 Yorum hurky

0004464 13 Mayıs 2014

İki Dakikada Bilim - Kas Tutulması

Ntv kanalında son zamanlarda reklam aralarında yayınlanan küçük, kısa bilgi alışverişi olan 2 dakikada bilim adlı program. Almanyada Ard kanalında 2008 yılında gösterime girmiştir.

62

37 Yorum duccio

0008019 24 Mayıs 2014

İki Dakikada Bilim - Vücudumuzdaki Uyuşmalar Neden Olur?

Ntv kanalında son zamanlarda reklam aralarında yayınlanan küçük, kısa bilgi alışverişi olan 2 dakikada bilim adlı program. Almanyada Ard kanalında 2008 yılında gösterime girmiştir.

59

24 Yorum 2dakikadabilim

0010699 24 Mayıs 2014

İki Dakikada Bilim - Tüylerimiz Neden Ürperir

Ntv kanalında son zamanlarda reklam aralarında yayınlanan küçük, kısa bilgi alışverişi olan 2 dakikada bilim adlı program. Almanyada Ard kanalında 2008 yılında gösterime girmiştir.

30

Figure 10: A Snapshot for Educational Videos

Appendix B: Visualization of Evaluation Criteria Results

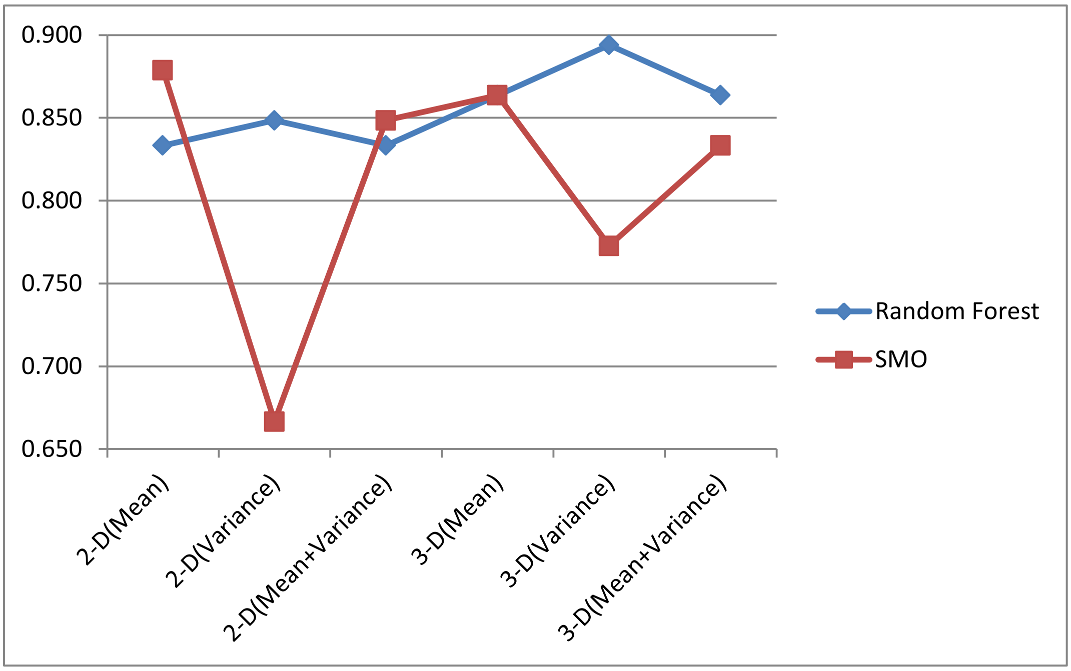


Figure 11: Accuracy Results of Features

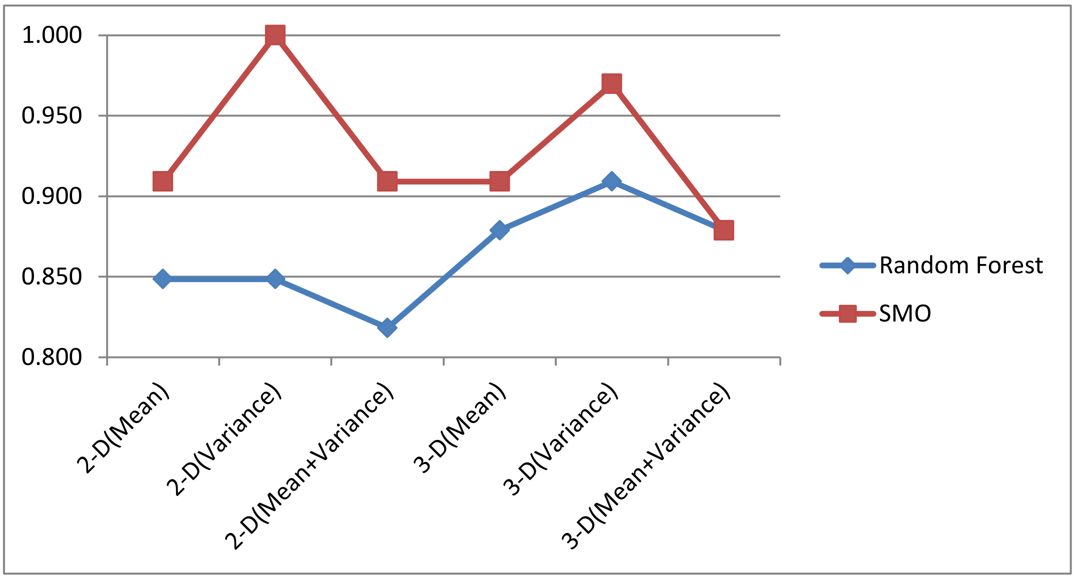


Figure 12: Recall Values In Terms of Features

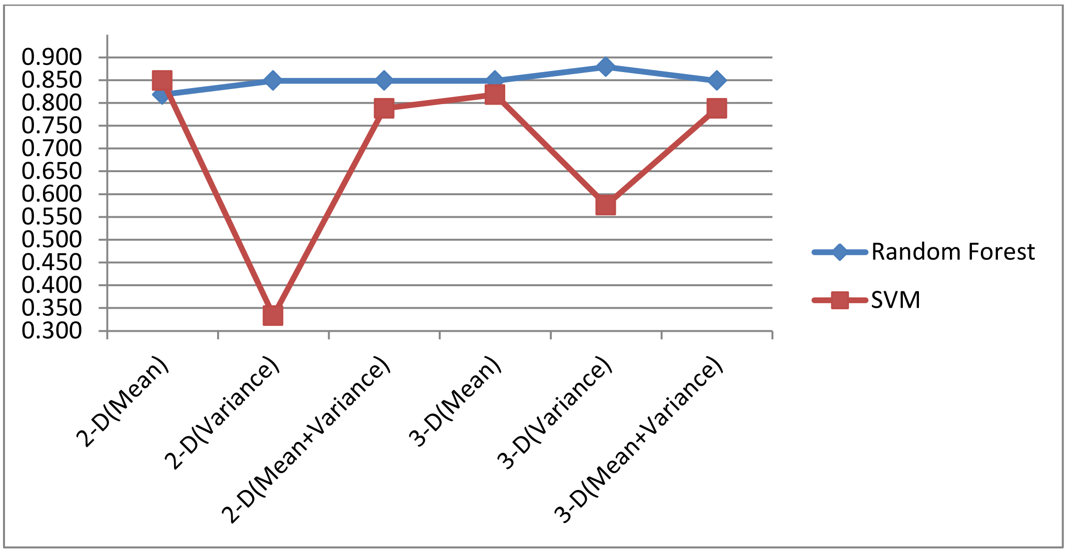


Figure 13: Specificity Values In Terms of Features

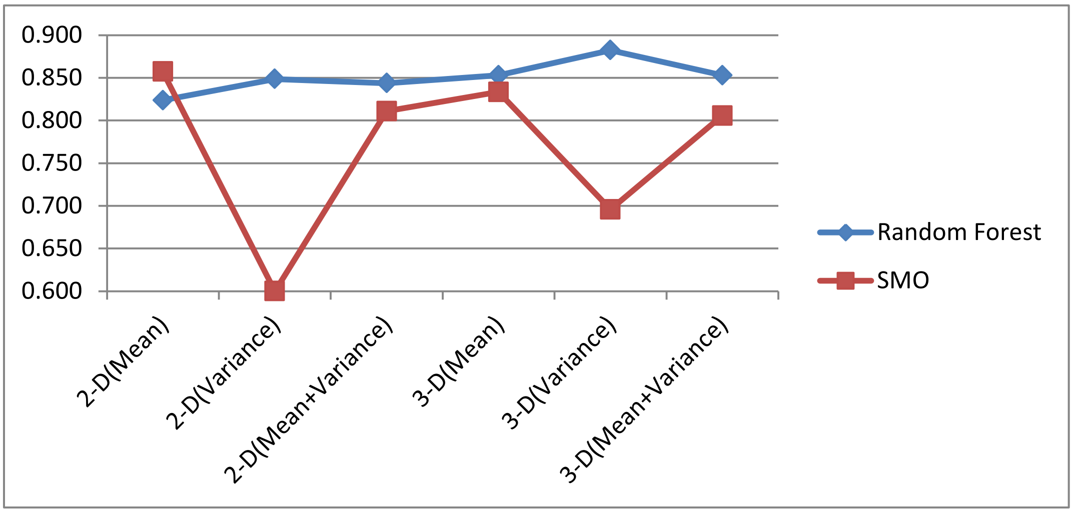


Figure 14: Precision Values In Terms of Features

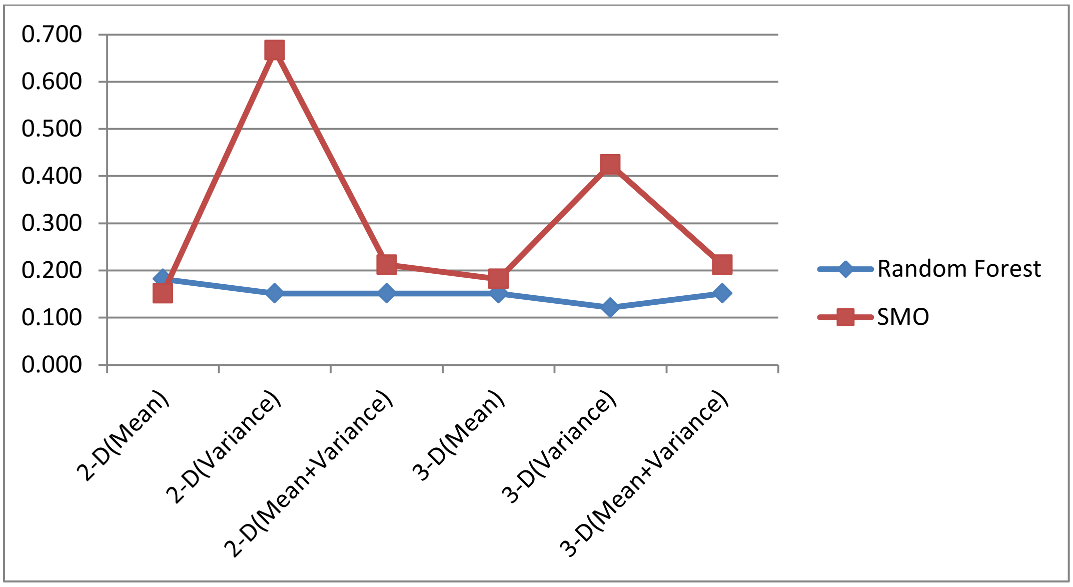


Figure 15: Fall-out Values In Terms of Features

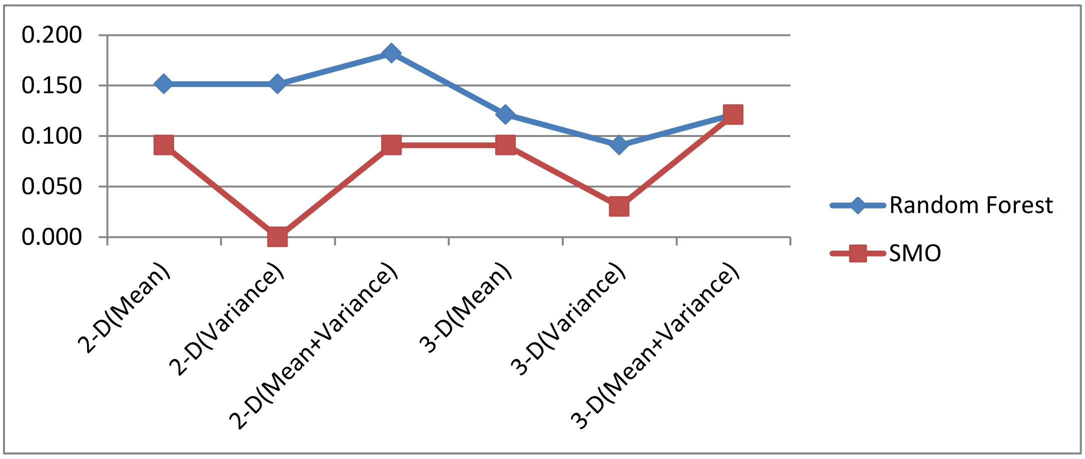
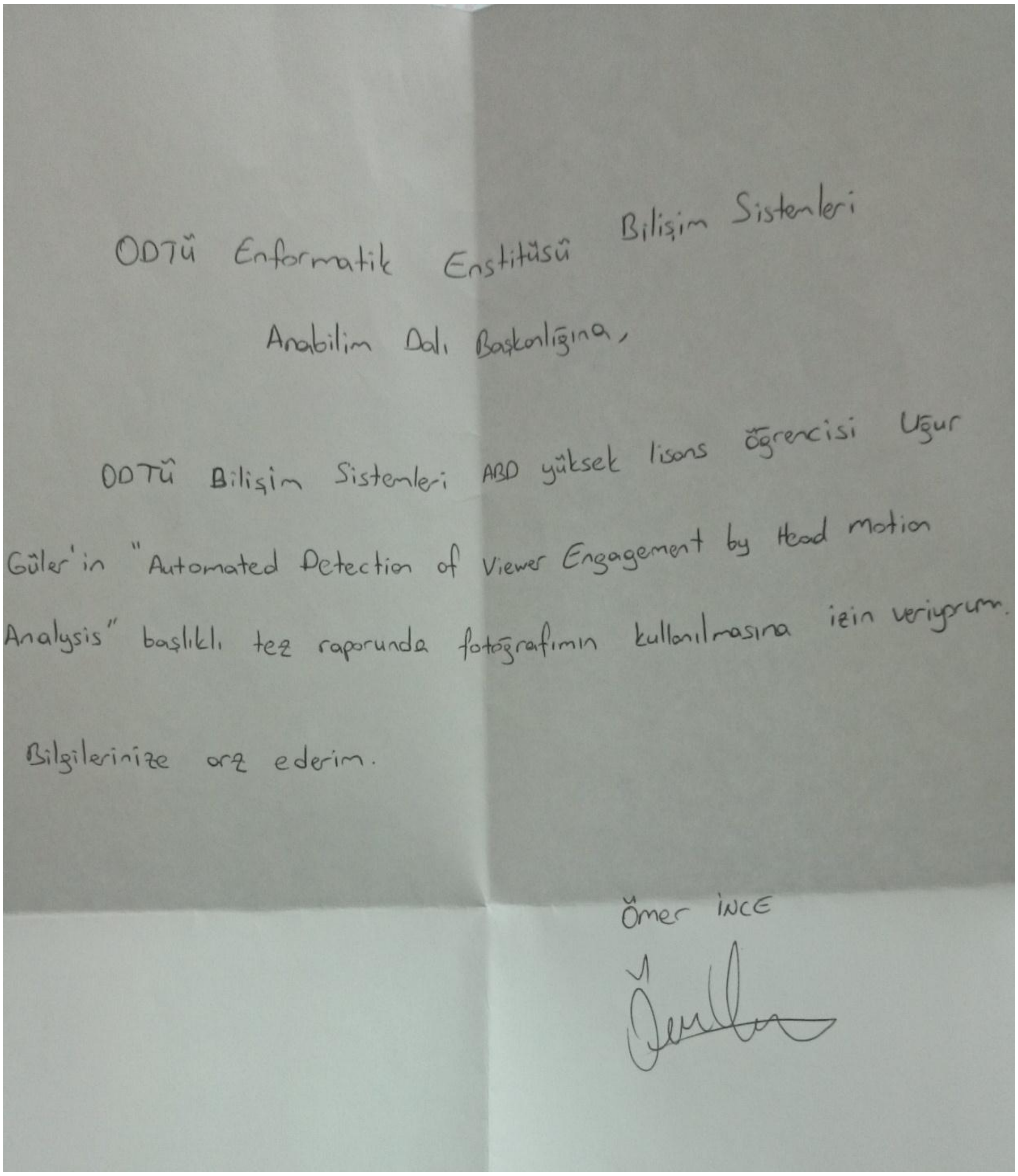


Figure 16: Miss-rate Values In Terms of Features

Appendix C: Utilization Permit



ODTÜ Enformatik Enstitüsü Bilişim Sistemleri

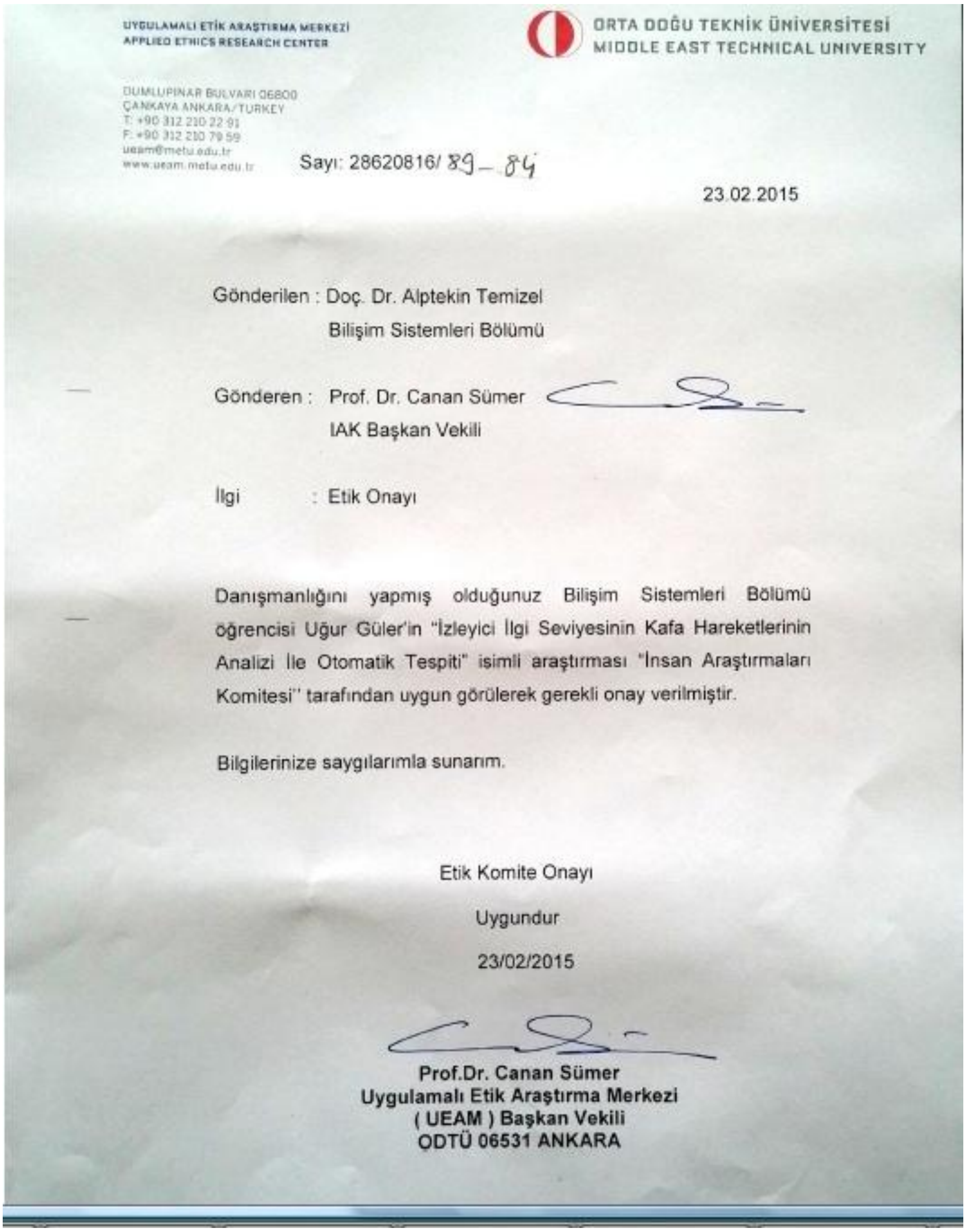
Anabilim Dalı Başkanlığına,

ODTÜ Bilişim Sistemleri ABD yüksek lisans öğrencisi Uğur Güler'in "Automated Detection of Viewer Engagement by Head motion Analysis" başlıklı tez raporunda fotoğrafımın kullanılmasına izin veriyorum.

Bilgilerinize arz ederim.

Ömer İNCE

Appendix D: Ethical Committee Approval



UYGULAMALI ETİK ARAŞTIRMA MERKEZİ
APPLIED ETHICS RESEARCH CENTER

ORTA DOĞU TEKNİK ÜNİVERSİTESİ
MIDDLE EAST TECHNICAL UNIVERSITY

DUMLUPINAR BULVARI 06800
ÇANKAYA ANKARA/TURKEY
T: +90 312 210 22 91
F: +90 312 210 79 59
ueam@metu.edu.tr
www.ueam.metu.edu.tr

Sayı: 28620816/ 89 – 84

23.02.2015

Gönderilen : Doç. Dr. Alptekin Temizel
Bilişim Sistemleri Bölümü

Gönderen : Prof. Dr. Canan Sümer
IAK Başkan Vekili

İlgi : Etik Onayı

Danışmanlığını yapmış olduğunuz Bilişim Sistemleri Bölümü öğrencisi Uğur Güler'in "İzleyici İlgi Seviyesinin Kafa Hareketlerinin Analizi İle Otomatik Tespiti" isimli araştırması "İnsan Araştırmaları Komitesi" tarafından uygun görülerek gerekli onay verilmiştir.

Bilgilerinize saygılarımla sunarım.

Etik Komite Onayı
Uygundur
23/02/2015

Prof.Dr. Canan Sümer
Uygulamalı Etik Araştırma Merkezi
(UEAM) Başkan Vekili
ODTÜ 06531 ANKARA

TEZ FOTOKOPİSİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü □

Sosyal Bilimler Enstitüsü □

Uygulamalı Matematik Enstitüsü □

Enformatik Enstitüsü X

Deniz Bilimleri Enstitüsü □

YAZARIN

Soyadı :Güler

Adı : Uğur

Bölümü :BİLİŞİM SİSTEMLERİ

TEZİN ADI (İngilizce) : AUTOMATED DETECTION OF VIEWER ENGAGEMENT BY HEAD MOTION ANALYSIS

TEZİN TÜRÜ : Yüksek Lisans X Doktora □

1. Tezimin tamamından kaynak gösterilmek şartıyla fotokopi alınabilir. X
2. Tezimin içindekiler sayfası, özet, indeks sayfalarından ve/veya bir bölümünden □
   Kaynak gösterilmek şartıyla fotokopi alınabilir.
3. Tezimden bir (1) yıl süreyle fotokopi alınamaz. □

TEZİN KÜTÜPHANEYE TESLİM TARİHİ :…………………….