OPTICAL FLOW BASED VIDEO FRAME SEGMENTATION AND SEGMENT CLASSIFICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

SAMET AKPINAR

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
COMPUTER ENGINEERING

FEBRUARY 2018

Approval of the thesis:

OPTICAL FLOW BASED VIDEO FRAME SEGMENTATION AND SEGMENT CLASSIFICATION

submitted by **SAMET AKPINAR** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Gülbin Dural Ünver _____
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. Halit Oğuztüzün _____
Head of Department, **Computer Engineering**

Prof. Dr. Ferda Nur Alpaslan _____
Supervisor, **Computer Engineering Department**

Examining Committee Members:

Prof. Dr. Uğur Halıcı _____
Electrical and Electronics Engineering Dept., METU

Prof. Dr. Ferda Nur Alpaslan _____
Computer Engineering Dept., METU

Prof. Dr. Mehmet Reşit Tolun _____
Computer Engineering Dept., Başkent University

Asst. Prof. Dr. Şeyda Ertekin _____
Computer Engineering Dept., METU

Asst. Prof. Dr. Orkunt Sabuncu _____
Computer Engineering Dept., TED University

Date: 01.02.2018

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 results that are not original to this work.

Name, Last name: Samet Akpınar

Signature:

ABSTRACT

OPTICAL FLOW BASED VIDEO FRAME SEGMENTATION AND SEGMENT CLASSIFICATION

Akpınar, Samet
Ph.D., Department of Computer Engineering
Supervisor: Prof. Dr. Ferda Nur Alpaslan

February 2018, 90 pages

Video information retrieval is a field of multimedia research enabling us to extract desired semantic information from video data. In content-based video information retrieval, visual content obtained from video scenes is utilized. For developing methods to cope with content-based video information retrieval in terms of temporal concepts such as action, event, etc., representation of temporal information becomes critical. In this thesis, action detection is tackled based on a temporal video representation model. Herein, the visual feature - optical flow - is our basic construct used to formalize video parts as temporal information. In the proposed model, video action detection is considered over a pieced approach composed of two parts;

Temporal video segment classification and temporal video segmentation. In the first part, weighted frame velocity concept is put forward and associated with the optical flow vectors. The associated representation is used in action based video segment classification. The second part contains a new temporal video segmentation methodology providing segment candidates to segment classification methods generally. The methodology brings an approach strengthening the pixel based cut detection methods with the motion based ones. Average motion vectors are presented based on the optical flow vectors and used in pixel matching. A binary cut classification is applied to the obtained representation enriched with a sliding window based approach. Proposed methods are applied to different data sets. Analysis of the results with the state of the art methods shows that proposed temporal representation models and concepts increased the segment and cut classification performances.

Keywords: Content-based video information retrieval, action detection, temporal video segment classification, temporal video segmentation, optical flow

ÖZ

OPTİK AKIŞ TABANLI VİDEO ÇERÇEVE BÖLÜMLENDİRME VE BÖLÜM SINIFLANDIRMA

Akpınar, Samet
Doktora, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Ferda Nur Alpaslan

Şubat 2018, 90 sayfa

Video bilgi getirme, çoklu ortam araştırmalarının, video verisinden istenen anlamlı bilgileri çıkarılmasını sağlayan bir alanıdır. İçerik tabanlı video bilgi getirmede, video sahnelerinden elde edilen görsel içerikler kullanılmaktadır. İçerik tabanlı video bilgi getirmeyi eylem, olay, vb. zamansal kavramlar açısından gerçekleştirmek için zamansal bilginin gösterimi kritik hale gelmektedir. Bu tezde, eylem tespiti, zamansal bir video gösterim modeline dayalı olarak ele alınmaktadır. Burada görsel özellik - optik akış - video kısımlarını zamansal bilgi olarak biçimlendirmek için kullandığımız temel yapımızdır. Önerilen modelde, video eylem tespiti, iki kısımdan oluşan parçalı bir yaklaşım üzerinden düşünülmektedir; Zamansal video bölüm sınıflandırma ve zamansal video bölümlendirme. İlk kısımda, ağırlıklandırılmış

görüntü karesi hızı kavramı öne sürülmekte ve optik akış vektörleriyle ilişkilendirilmektedir. Birleştirilmiş bu gösterim, eylem tabanlı video bölüm sınıflandırmada kullanılmaktadır. İkinci kısım, genel olarak video bölüm sınıflandırma metotlarına video bölüm adaylarını sağlayan yeni bir zamansal video bölümlendirme metodolojisi içermektedir. Bu metodoloji, piksel tabanlı kesme tespit yöntemlerini, hareket tabanlı olanlarla güçlendiren bir yaklaşım getirmektedir. Optik akış vektörlerine dayalı olarak ortalama hareket vektörleri sunulmakta ve piksel eşlemede kullanılmaktadır. Bir kesme sınıflandırma, elde edilen kayan pencere tabanlı yaklaşımla zenginleştirilen gösterime uygulanmaktadır. Önerilen metotlar farklı veri kümelerine uygulanmıştır. Sonuçların literatürdeki metotlarla analizi, sunulan zamansal gösterim ve kavramların kısım ve kesme sınıflandırma performanslarını arttırdığını göstermektedir.

Anahtar kelimeler: İçerik tabanlı video bilgi getirme, eylem tespiti, zamansal video bölüm sınıflandırma, zamansal video bölümlendirme, optik akış

To my dearest son Saygın,

ACKNOWLEDGMENTS

I am sincerely grateful to my supervisor Prof. Dr. Ferda Nur Alpaslan for support, guidance, and encouragements during the thesis work. Her advices and endless efforts made it possible to overcome difficulties throughout the research.

I would like to express my deepest gratitude to my thesis examining committee members. I would like to thank Prof. Dr. Uğur Halıcı for her suggestions and comments and Asst. Prof. Dr. Şeyda Ertekin for her support and interest in this study.

It was a great pleasure to listen the suggestions and ideas of Prof. Dr. Nihan Kesim Çiçekli.

Also, special thanks to my family for their support and positive approach. Their patience and confidence have always provided great motivation and given great responsibility to me.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Information sharing platforms' becoming widely-spread under the influence of internet technology trends increased the needs for gathering required visual information from various sources in reasonable period of times. This exposed the necessities about structuring and interpreting the visual information in order to make it available for search. Regarding the tremendous amount of video archives, performing information extraction and retrieval becomes important for efficient search. Video information extraction and retrieval comprises research areas such as automatic video annotation, video action detection, and so on in order to make video information structured and semantically meaningful.

Video information retrieval is a field of multimedia research enabling us to extract and search the desired semantic information from video data having visual, audio and textual features. While textual features include high level semantic information, they lack the automatedness. The retrieval based on textual features strongly depends on the textual sources which are commonly created manually. On the other hand, audio features are restricted to a supervisor role. As the audio information does not have the structure expressing the video situations such as objects, actions, etc., generally, it can be used as an additional resource supporting visual and textual information. Visual video features provide the basic information for the video events or actions. Although it is difficult to obtain high levels of semantics by using visual information, a convincing way to construct an independent fully automated video information retrieval model is to utilize visual information as the central resource.

Content-based video information retrieval is the automatic or semi-automatic retrieval of conceptual video items such as object, action, event etc. using the visual content obtained from video frames. Visual features play critical roles at this point. There are many algorithms extracting visual features and using them for modeling retrieval methods. Their visual feature sets vary from static image features - pixel values, color histograms, edge histograms, and so on - to temporal visual features - interest point flows, shape descriptors, motion descriptors, etc. While static image features are more related with static concepts such as visual objects, temporal visual features are meaningful for temporal concepts such as video actions, events, trajectories, etc. Extraction, recognition and detection of these concepts are important ways of abstraction covered by content-based video information retrieval.

1.1 Problem Definition

In this thesis, the study is focused on one of the temporal concepts - video action - of content-based video information retrieval. Temporal visual features are the descriptors combining the visual image features with the time information. In this context, representing video information using temporal visual features generically means modeling the visual video information with temporal dimension. Accordingly, this modeling constructs the temporal video information.

The construction and representation of temporal video information complete each other. While representation utilizes the features of the constructed model, features and models used in the construction reflects the underlying elements of the representation formalism. Herein, construction and representation of temporal visual video information is together conceptualized as the representation of temporal video information.

For developing methods to cope with content-based video information retrieval effectively in terms of temporal concepts - action, event, trajectory, etc. -, we need to formalize the representation of temporal video information. Therefore, the main

problem can be reduced to representing temporal information. The purpose of the thesis is to solve the representation of temporal information problem in video action detection of content-based video information retrieval domain. This solution yields to a common representation formalism opening new perspectives for video action detection.

How to utilize the representation for action detection is one of the key points. While some action detection approaches provide unique processes targeting the extraction of video items directly from the videos, the others follow pieced mechanisms. Action detection, here, is considered as a pieced process composed of temporal video segmentation, that is frame segmentation, and temporal video segment classification, mainly segment classification. In order to classify the video scenes as video segments identified by actions, some sort of temporal segmentation is needed. Namely, temporal video segmentation can prepare the candidate segments for temporal video segment classification methods.

In both temporal video segment classification and temporal video segmentation, novel methods using the mentioned representation are needed. The study and contributions are considered in these two parts. Success of the proposed methods constitutes a contribution for action detection research as a whole.

1.2 Overview

In this thesis, a new model is proposed for video action detection. Visual features such as edges, visual interest points etc. of video frames are the basics for constructing the model. These features are used for constituting a more complicated motion feature, namely optical flow, which is the building block feature for the representation of temporal video information in our study.

In the proposed model, the approach spans through two dimensions - temporal video segment classification and temporal video segmentation. In this context, a temporal

video representation is proposed to formalize the video parts in defined time intervals as temporal information. The representation, which is determined to be generic for both temporal video segment classification and temporal video segmentation, is fundamentally based on the optical flow vectors calculated for the frequently selected frames of the video parts.

In the first dimension, a new temporal video segment classification method is proposed. Segments are intended to be classified according to defined actions. In the method, video segments are represented temporally based on the mentioned temporal video representation. Weighted frame velocity concept is put forward for a whole video segment in addition to the generic representation containing optical flow vectors, optical flow histograms and optical flow sums. Weighted frame velocity is combined with optical flow based entities according to proposed relations. The combined representation is used in the action based temporal video segment classification.

In the second dimension, a new temporal video segmentation method is modeled. Segments are tried to be obtained by detecting the cuts between the scenes. The methodology brings an approach strengthening the pixel based cut detection methods with the motion based ones. Calculated optical flow vectors discussed in the first dimension are also used here as building block features to represent scene changes. Namely, the generic temporal information representation approach is also utilized here. Average motion vectors are calculated based on optical flow vectors and also pixel matching based on these average vectors are made between consecutive frames. Pixel matching for each frame transition forms the representation formalism and its set. Then, the problem converges to binary cut classification. Binary cut classification works on this constructed set. The results are improved using specific information extraction methods.

Methods proposed in both temporal video segment classification and temporal video segmentation are applied to significant data sets and the results are analyzed by

comparing to the state-of-the art methods. Then, the contribution of the proposed approaches for temporal information representation in temporal video segment classification and temporal video segmentation are stated evidently. The study is organized by starting from representation, continuing with segment classification and segmentation.

1.3 Contribution

The thesis proposes a combined model using a temporal video representation for video action detection. Contributions of the thesis are summarized below.

First contribution is the proposed ***temporal video representation model***, which is a generic framework for both temporal video segment classification and temporal video segmentation. The representation is based on optical flow concept.

Second main contribution resides in the adaptation of the above mentioned representation to temporal video segment classification. The adaptation is based on a common way of partitioning optical flow vectors according to their angular features. An angular grouping of optical flow vectors is used for each frame of the video. We describe ***Weighted Frame Velocity*** concept, measuring a kind of velocity of the cumulative angular grouping of a temporal video segment in order to represent the motion of the frames of the segments more descriptively. The segment representation is utilized for classifying video segments as defined actions.

Last contribution contains the temporal video segmentation methodology combining a pixel based cut detection approach with a motion based approach. It also proposes an adaptation of the generic temporal representation. A spatial partitioning of optical flow vectors is used. The pixel matching algorithm is defined by using optical flow vectors and newly proposed ***Average Motion Vector*** concept, calculated from optical flow vectors for each spatial group. The algorithm tries to classify the cuts using block distances of spatial groups calculated from pixel matching between two

consecutive selected frames. The results are improved using a ***Sliding Windows Method***.

Thus, the new optical flow based representations and methods for both temporal segment classification and temporal video segmentation are the main contributions of this thesis in action detection research. In temporal segment classification, the contribution of the new optical flow based representation to the standard optical flow based approaches and the advantage over the bag-of-words approaches is shown. On the other hand, the success of the representation and method in temporal video segmentation compared to threshold based, histogram based and standard optical flow based methods is presented.

1.4 Outline

This chapter is the introduction. In Chapter 2, related work about content-based video information retrieval, representation of temporal video parts, temporal video segment classification and temporal video segmentation is proposed. Accordingly, Chapter 3 is about Optical Flow Based Temporal Representation, Chapter 4 explains Temporal Video Segment Classification and Chapter 5 contains Temporal Video Segmentation. In Chapter 6, Experiments, Results are discussed, and Improvements are introduced. Last, Conclusion and Future Work is given in Chapter 7.

CHAPTER 2

RELATED WORK

Video information retrieval is the process of automatic annotation and retrieval of video content. Various types of data sources such as textual, visual, audio, and so on are used in video information retrieval. In this study, the focus is on temporal visual features of videos. Therefore, related work is organized accordingly.

2.1 Content-Based Video Information Retrieval and Action Detection

We narrowed down our study to content-based video information retrieval which deals with the visual content and concepts of video information for retrieval. Action detection is our specific problem domain in content-based information retrieval.

In [9], realistic human actions are tried to be annotated temporally and automatically in videos. The available manual annotations are used as the support information. In this context, the study is focused on both learning the actions from the available manual annotations and localizing them. Namely, textual and visual features are utilized together. Movie scripts are used for getting rid of the training cost of manual annotation. But, regarding the nature of movie scripts, precise information about the actions - especially their location - cannot be obtained. In order to solve the problem, they propose a kernel-based discriminative clustering algorithm carrying out the localization of actions. According to the localized action candidates, temporal action detectors are designed and trained. Namely, content-based video information retrieval is supervised by textual information for localization.

[38] introduces a new method for human action recognition. The method contains a model splitting the foreground and background motion and actions. Then, the actions are recognized using the new representation based on this model.

The study of [55] mainly contains an event detection method using a segment-based approach for motion features. The method is originated from the idea that the motion features are critical for detecting events because events may have some specific actions or motion patterns. A segment-based video representation is proposed. Videos are represented as segments for feature extraction and classification.

[56] tackles human activity recognition using optical flow based features. An optical flow based approach for recognizing human actions in video sequences is shown. A visual descriptor using optical flow vectors along the edges is proposed. A multi-class SVM classifier is applied to the feature vectors for activity recognition.

In [60], an algorithm detecting abnormal events is proposed. The algorithm utilizes Histograms of the Orientation of Optical Flow (HOOF) as the motion descriptor. One class SVM classifier is applied. Grid concept is introduced for HOOF. It behaves as a descriptor for the motion information of the monolithic video frame. SVM - one class - detects the abnormality in the current frame in order to label it as the abnormal event.

[78] tackles action recognition by trying to find frame regions which are specific to certain actions. The proposed method extracts the spatial regions by applying a matrix factorization on optical flow fields. Consequently, action recognition is carried out by characterizing the distribution of extracted regions.

[79] presents an action recognition method based on deep convolutional neural networks. The method combines different sources of information for deep learning. The spatio-temporal features obtained from the constructed spatial network are

amplified by using optical flow vectors. It increases the performance of deep convolutional neural network.

In [80], a novel method is proposed for learning spatio-temporal features automatically for action recognition. A genetic programming approach is used for extracting features from color and optical flow fields.

2.2 Representation of Video Parts Temporally

Temporal representation of video parts is critical in various areas of content-based video information retrieval. Video action detection, video action recognition, event detection, object tracking, and video motion analysis are some prominent issues which exemplifies this criticality.

There are different approaches in the representation of temporal video parts. The studies in [10, 11, 12 and 13] focus on the perception of the visual world and their studies bring us facts about how to detect the visual features in which context more philosophically. Regarding the visual features, mentioned approaches can generally be figured out. The methods can be grouped according to these approaches.

Key-frame, bag-of-words, interest points and motion based approaches are the groups reflecting the way of representation.

Key-frame based representation approaches focus on detecting key frames in the video segments in order to use them in classification. This kind of representation is used in [1, 2, 3 and 14] for video scene detection and video summarization. The study of [1] contains the segmentation of videos into shots and key-frames are extracted from these shots. As the scene duration is not known, first, the shots are assigned to visual similarity groups. Then, each shot is labeled according to its group and a sequence alignment algorithm is applied for recognizing the shot labels' change patterns. Shot similarity is estimated using visual features and shot orders are

kept while applying sequence alignment. In [2], a novel method for automatic annotation of images and videos is presented with keywords among the words representing the concepts needed in content-based image retrieval. Key-frame based approach is used for videos. Images are represented as the vectors of feature vectors containing visual features such as color, edge, etc. They are modeled by a Hidden Markov Model (HMM); whose states represent concepts. Model parameters are estimated from a training set. The study proposed in [3] deals with automatic annotation and retrieval for videos using key frames. They propose some new approach automatically annotating video shots with semantic concepts. Then, the retrieval carried out by textual queries. An efficient method extracting Semantic Candidate Set (SCS) of video shots is presented based on key-frames. Extraction uses visual features. In [14], an innovative algorithm for key frame extraction is proposed. The method is used for video summarization. Metrics are proposed for measuring the quality.

Histogram based bag-of-words (BoW) approaches represent the frames of the video segments over a vocabulary of visual features. [4 and 5] are the examples of such approaches. [4] proposes a method interpreting temporal information with the BoW approach. Video events are conceptualized as vectors composed of histograms of visual features, extracted from the video frames using BoW model. The vectors, in fact, can be behaved as the sequences, like strings, in which histograms are considered as characters. Classification of these sequences having difference in length, depending on the video scene length, is carried out by using Support Vector Machines (SVM). SVM has a string kernel based on the Needleman-Wunsch edit distance. In [5], a new motion feature is proposed, Expanded Relative Motion Histogram of Bag-of-Visual-Words (ERMH-BoW) implementing motion relativity and visual relatedness needed in event detection. Concerning the ERMH-BoW feature, relative motion histograms are formed between visual words representing the object activities and events.

Despite their performance issues in terms of time complexity, above approaches lack the flow features and temporal semantics of motion although they are efficient in spatial level. On the other hand, motion based approaches deal with motion features which are important in terms of their strong information content and stability over spatio-temporal visual changes. Motion features such as interest points, optical flow, etc. are used for modeling temporal video segments. [6, 7, 15 and 35] are the studies using motion features. [6] proposes new framework in order to group the similar shots into one scene. Motion characterization and background segmentation are the most important concepts in this study. Motion characterization results in the video representation formalism while background segmentation provides the background reconstruction which is integrated to scene change detection. These two concepts and the color histogram intersection together become the fundamental approach for calculating the similarity of scenes. The study of [7], presents a new approach which implements motion estimation in video scenes. The representation of video motion is carried out by using some sort of particles. Herein, particles are image points with their features and trajectories. Appearance stability along the particle trajectories and defects between the particles are measured for optimizing the trajectories. The motion representation can be used in various areas. It cannot be constructed using the standard methods such as optical flow, feature tracking, etc. Optical flow is a spatio-temporal motion feature describing the motion of visual features. Optical flow based representation is especially strong for video segment classification. [15, 35 and 72] present methods for representing video segments with optical flow. [15] proposes a representation structure based on direction histograms of optical flow. In [35], video segments are tried to be represented by using a histogram of oriented optical flow (HOOF). By the help of this representation, human actions are recognized by classifying HOOF time-series. A generalization of the Binet-Cauchy kernels to nonlinear dynamical systems (NLDS) is proposed for classification. The study of [72] brings a new representation based on dense trajectory features. It is enriched with camera motion estimation. Namely, feature point matches between frames using SURF descriptors and dense optical flow are extracted. The resulting representation is a standard bag-of-words (BOW) histogram.

2.3 Temporal Video Segment Classification

Temporal video segment classification is an important sub-problem in content-based video information retrieval. By definition, it is the classification of scenes in a video. The classification highly depends on the representation of temporal video information and the classification methods working on this representation.

[34, 37, 42 and 43] propose the approaches based on 3D interest points. These methods tackle the problem of video segment classification by putting new interest points or visual features forward by enriching with time dimension. Therefore, the features in the studies can be conceptualized as space-time shapes.

The methods introduced in [44 and 45] views the problem from the point of spatio-temporal words. The segments are bag-of-features and make the classification according to the code words.

[63] proposes a method for action recognition using the trajectories of selected features. Trajectory fragments of tracked features are extracted and valued. The fragments and their values are based on feature tracking.

[15 and 35] present optical flow based methods for video segment classification. Optical flow histograms are constructed and utilized in segment representation. Then, this representation is used for classifying the segments.

In [59], the study is focused on the integration of computer vision approaches to surveillance video analysis. The main purpose is to extract the video segments symbolizing human actions. Histogram of oriented optical flow (HOOF) is used for region based activity recognition among the other techniques.

The studies proposed in [61, 62, 64, 65, 66, 67 and 71] utilize optical flow as an important visual feature in order to carry out action recognition. [61] improves the performance of dense trajectories - as an efficient video representation - by

supervising with camera motion. Feature points between frames are matched by SURF descriptors and dense optical flow. These matches are used to find a homography to remove the camera motion from the optical flow. This is exactly an improvement for motion-based descriptors, such as HOF and MBH. In [62], a novel approach to action recognition using optical flow analysis is explained. Useful properties of optical flow are determined by a covariance matrix of flow velocity, gradient and divergence. [64] presents an effective motion descriptor based on HOOF. Principal Component Analysis (PCA) is applied for dimensional reduction. Then, action recognition is done by using the hidden Markov model (HMM). [65] proposes a method for human action recognition based on motion patterns. Motion features constructed using optical flow vectors are used. The features inspecting local regions of the image sequence and are created by a variant of *AdaBoost*. The features are tuned to determine different classes of action, and have better time complexity. [66] put forward some optical flow based kinematic features containing divergence, vorticity, symmetric/antisymmetric flow fields, second/third principal invariants of flow gradient, and third principal invariant of rate of rotation tensor. In this context, kinematic features represent spatiotemporal patterns having the ability to reflect the representative dynamics of the optical flow. Principal component analysis (PCA) is applied on the kinematic features. Last, multiple instance learning (MIL) is used for classification. Video actions are represented by a bag of kinematic labels in the classification. In the study [71], human actions are tried to be recognized using shape and optical flow features together. [68, 69 and 70] are also proposing methods mainly utilized for action recognition or classification. Again, optical flow features play critical roles alongside the other visual features. One of the common properties of all these optical flows based studies is that the representation of video segments is enriched with many motion features and a combined representation is put forward originally.

2.4 Temporal Video Segmentation

Temporal video segmentation is another dimension in content-based video information retrieval methods. It is defined as extracting the scenes from a video instance. This can be done by using different approaches. While the extraction can be done by behaving the whole video information totally, on the other hand, it can be done by detecting the cut points.

The scheme defined in [48] suggests a difference based cut detection method using threshold detection. The method is based on quantifying the interframe differences. A metric for dissimilarity is proposed and an automatic threshold detection algorithm based on feature tracking is created. According to the detected threshold, a frame's being cut or noncut is determined.

In [46 and 47], methods are proposed based on novel complicated visual based features enabling cut classification without using thresholding. [46] comes up with new features which are specifically designed to detect the differences between cuts and normal sequences. The features are based on normalized color histogram. A difference measure for the histograms is specifically created as to be suitable for separating cut and normal shots. At the end, feature vectors are created with the calculation of the measure for each consecutive frame pair. SVM is used for classifying the feature vectors as cut or normal shot. In [47], tridimensional image feature vectors, which are extracted using 2D Gabor filtering, are designed. An unsupervised learning procedure is defined instead of thresholding. A region growing based approach on the distances between the consecutive feature vectors is provided. Resulting binary space is the result of shot detection.

In the study [58], a new concept - joint histogram - is put forward in order to extend the color histogram which has great disadvantages in visual similarity. As the histogram similarity does not always mean the real similarity in terms of visual content, joint histogram is suggested as a solution. Joint histogram is a

multidimensional histogram of visual features such as color, edge density, texturedness, gradient magnitude and pixel rank. The histogram is composed of pixel counts for each combination of selected feature dimensions. Different image histograms are described by different combinations of features and their performance is evaluated. Although the study is focused on comparing images, the concept is also useful for detecting cut frames.

[49] proposes a cut detection method using fuzzy rules. These rules are applied to the representation of video frames using spatio-temporal features.

Cut detection methods based on optical flow features are presented in [39, 40, 41, 51, 52, 53 and 54]. The study [39] proposes an optical flow based model using linear prediction. Cut detection is recognized in various transition conditions. Video frames are divided into 4×4 non-overlapping blocks. The proposed method reduces the cut detection problem into a backward block searching for a match. Three consecutive frames with the same matched block and the fourth predicted frame are grouped as a subsequence. [40] presents motion based methods using distribution of optical flow in terms of spatial locations. [41] introduces a novel method for temporal segmentation of a video based on motion features. Markov Random Field is constructed for ensuring temporal continuity. Optical flow is the key motion feature used in the method. The approach is based on the fact that, optical flow changes differ at shot boundary points. In [51], a scene change detection algorithm is presented. It works as trying to find candidate video frames that may include scene change moments. The novel model exploits optical flow in terms of its statistical features. First, optical flows are partitioned into background and foreground groups. Afterwards, scene changes between consecutive frames are analyzed by using the optical flow and statistical methods, such as average and standard deviation are applied to evaluate the scene change probability of the frames. [52] proposes a learning-based approach for motion boundary detection. The main idea, here, is “discontinuities of the optical flow field corresponds to motion boundaries “. A

structured random forest is used for predicting motion boundaries. The random forest leverages mainly color values and optical flow vectors.

In [53 and 54], occlusion boundary detection is tackled. [53] makes use of optical flow to create a contour and region detector which distinguishes occlusion boundaries from internal boundaries. In [54], occlusion boundary detection is improved via enhanced exploration of contextual information including optical flow. The method tries to find the optical flow discontinuity. A novel approach based on convolutional neural networks (CNNs) and conditional random fields (CRFs) is proposed.

According to the above studies, content-based video information retrieval, action detection, representation of video parts temporally, temporal video segment classification and temporal video segmentation problems can be viewed more definitely. The approaches tackling these problems will be shaped in the appropriate fields discussed so far.

CHAPTER 3

OPTICAL FLOW BASED TEMPORAL REPRESENTATION

Temporal video representation is the problem of representing video information - scene, segment, etc. - as meaningful temporal descriptors which express the video parts in defined time intervals. While this problem generally runs through the video information including visual, audio and textual features, this thesis deals with visual features only. Thus, temporal visual video concept will be called as temporal video concept in this context.

The above mentioned research problem is originated from representing the temporal information. Temporal information contains time and magnitude for a logical or physical entity. Robot sensor data, web logs, weather, video motion and network flows are common examples of temporal information. Independent from domain, both representation and processing methods of temporal information is important in the resulting models. Regarding the processing methods, prediction, classification and mining can be considered as first comers for the temporal information. In most cases, the representation is also a part of the processing methods. While the representation and processing methods are handled together, the focus is especially on the processing methods rather than the representation in most cases, such as temporal data mining and time series classification.

The types of the features and their quality on describing the domain knowledge also influence the temporal information processing. Also, having high dimensionality makes the effective representation of temporal information with more complicated

features important. Therefore, feature definitions, construction and feature extraction methods play an important role for processing the temporal information. This point of view can contribute surprisingly in many studies, as it utilizes domain knowledge with features in one sense. As the focus, here, is feature extraction and construction, the improvements are measured with common methods. In classification, for instance, the information is represented by newly generated features, the data model is generated for the target machine learning method and the contribution of newly proposed features is measured by classifying the modeled information with the related method.

In content-based video information retrieval, visual video data behaves as a kind of temporal information including frame sequences along the time. Each frame of the video has its visual information along with its time value. The temporal information representation highly depends on the visual content of video frames. The basic and most primitive representation of temporal video information is the representation of the all pixel intensities of all frames of the video. While this representation includes the richest visual information, processing and interpreting this information is impossible. In a 600 x 480 frame size for a 10 seconds scene - 30 fpsec, 86.4M features exist with this approach. Therefore, there is a need for efficient representation formalisms.

Key-frame based representation is one of the candidate approaches for representing temporal information in videos. For each scene, a key-frame is selected according to some calculations on visual features. By using this key frame, whole scene is represented and feature size for the representation is decreased. But, there is an important problem in key-frame based approaches. A key-frame including a plane is shown in Figure 3.1.

Figure 3.1: A key frame for a plane action

As we see here, it is obvious that a plane is flying in this frame. But, it is impossible to comment on the action of the plane as it may take off or land. Then, key-frame based approaches lack the important information resulting from the motion in the videos. Moreover, ignoring all of the frames except the key-frames makes the approaches mostly inapplicable for detecting boundaries in temporal video segmentation problems.

Another approach is bag-of-words approach for frame sequences. In this kind of representation, frames are behaved as code words obtained from grouping of the frames according to the visual features. With these code words, frame sequences are represented as sentences. This kind of representation includes the temporal nature of the scenes. But, the most important disadvantage of this representation style is the restricted nature of code words. Representing a visually rich frame with a label means lacking lots of information. The representation is restricted with the variety of the code words. Therefore, limitless types of frames will be reduced to very limited number of labels.

“Interest points” based representation is alternative representation formalism for temporal video information. Interest points are the “important” features that may best

represent the video frames invariant from the scale and noise. This representation alternative is very successful in reducing the huge frame information into small but descriptive patterns. But, it is again disadvantageous in detecting motion features despite its descriptiveness. As the motion features include flow with time, it is important to track the features along the time. Using interest points for representation, then, lack the motion based information.

State-space approaches are also a candidate for representing temporal video information. State-space methods define features which span the time. Space-time interest point concept proposed by [34] is an excellent example for state-space methods. Interest points which are spatially defined and extracted in 2D are extended with time. With this extension, interest points gain a 3D structure with time. Therefore, a space-time 3D sketch of frame patterns can be obtained, and they are ready for processing. State-space approaches best fit the representation of video information temporally as they can associate the time with the visual information in a descriptive and integrated way.

In this thesis, a state-space based representation approach is selected. Optical flow is the motion feature (integrating time with visual features) utilized for constituting the state-space method.

3.1 Optical Flow

Theoretically, optical flow is the motion of visual features such as points, objects, shapes etc. through a continuous view of the environment. It represents motion of the environment relative to an observer. James Jerome Gibson firstly introduced the optical flow concept in 1940s, during World War II [16]. He was working on pilot selection, training and testing. He intended to train the perception of pilots during the war. Perception was considered for the effect of the motion on the observer. In this context, shape of objects, movement of entities, etc. are handled for perception. During his study on aviation, he discovered optical flow patterns. He found that the

environment observed by the pilot tends to move away from the landing point, while the landing point does not move according to the pilot. Therefore, he joined this concept with the pilot perception on the observed environment. In Figure 3.2 [16], landing plane is shown with optical flow departing from the landing point using the pilot view.

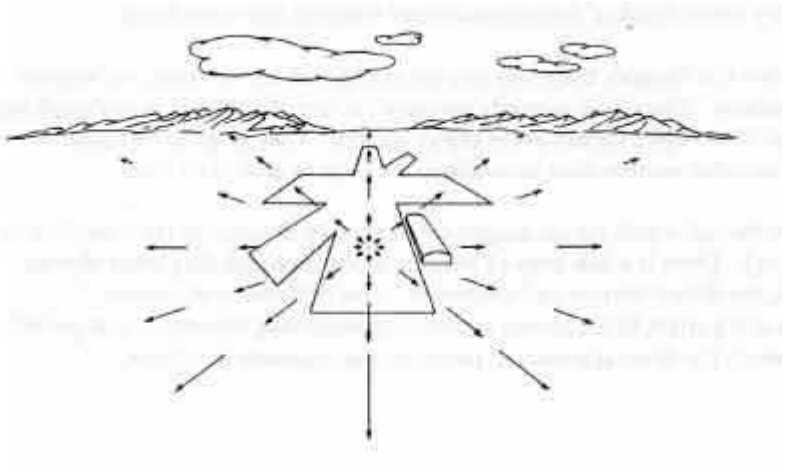


Figure 3.2: Landing plane with optical flow

In the perception of an observer, there may be two options for approaching/departing optical flow around a point. In the first option, the observer may be moving through the target point. This makes the optical flow departing from the point. In the second option, the environment around the point may be moving through the motionless observer. This also gives the same effect, having the optical flow departing from the point. These two options are also valid for approaching optical flow. If the observer departs from the target point or the point departs from the motionless observer, the optical flow is seen as approaching through the point. In Figure 3.3 [16], a moving train is shown with optical flow approaching to the observed point from the view of an observer looking from the back of the train.

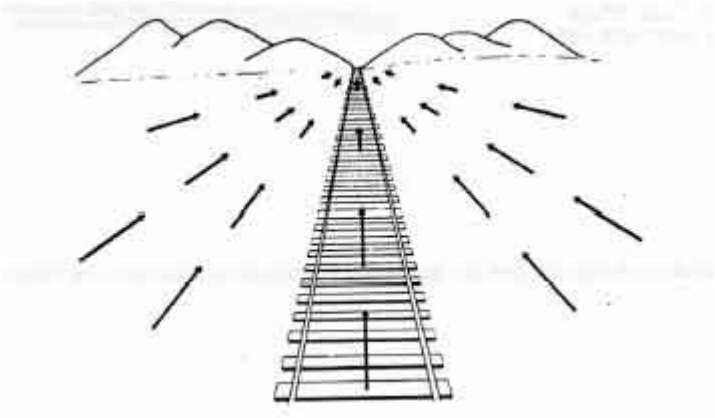


Figure 3.3: Departing train with optical flow

Optical flow is defined as the apparent motion of brightness patterns in the images in video domain [76]. More specifically, it can be conceptualized as the motion of visual features such as corners, edges, ridges, textures etc. through the consecutive frames of a video scene. Optical flow, here, is materialized by optical flow vectors. An optical flow vector is defined for a point (pixel) of a video frame. In optical flow estimation of a video frame, selection of "descriptive" points is important. This selection is done using visual features. It is clear that using an edge point or corner point is more informative than using an ordinary point semantically as the motion perception of human is based on prominent entities instead of ordinary ones. Optical flow vectors are, then, the optical flow of video frame feature instances instead of all frame points.

In brief, two problems arise in the optical flow estimation of video frames:

- Detection and extraction of the features to be tracked
- Calculation of the optical flow vectors of the extracted features

Optical flow estimation is finding effective solutions to these problems. Calculation of optical flow vectors of the extracted features can be reduced to the following problem;

"Given a set of points in a video frame, finding the same points in another frame"

In this aspect, detection and extraction of features depend on the needs of optical flow calculation. The goal is to find features whose motion can be tracked soundly.

3.1.1 Optical Flow Estimation

There are various approaches concerning the estimation of optical flow. *Differential, region-based, energy-based, phased-based* methods are the main groups of approaches dealing with the estimation of optical flow [17]. All of these groups include many algorithms proposed so far. Each of these algorithms reflects the theoretical background of its group of approach.

Here, the meaning of optical flow estimation is discussed from a *differential* point of view. The explanation is based on the change of pixels with respect to time. The solution of the problem can be reduced to the solution of the following equation [19]:

$$I(x + \delta x, y + \delta y, t + \delta t) = I(x, y, t) \quad (1)$$

The equation is written for a point in a video frame. The point is assumed to change its pixel value over time. (x, y, t) is defined as the 3D point composed of the 2D coordinates (pixel location) of the given point at time t. I represents the intensity function giving the image intensity value of a given pixel value at a given time. $\delta x, \delta y$ are the amount of location changes in x and y directions and δt is the change of time between two frames. The equation is based on the assumption in which enormously small amount of change on the pixel position of the point in enormously small amount of time period converges to zero change in the intensity value. Another

words, it says that, the intensity value of a pixel in a frame is equal to the intensity value of another pixel having the same point - the point in the former pixel - in the next frame. The point moves enormously small amount of distance - pixel change - in enormously small amount of time. Herein, Taylor Series plays a crucial role.

Taylor Series

A Taylor Series [73, 74 and 75] is another form of a function as an infinite sum of terms obtained from derivatives of the function at some point. Theoretically, Taylor Series representation has infinite number of terms. But, function approximations can be done with the finite number of terms of Taylor Series. The number of terms used in the Taylor Series determines the error of the approximation. The order of Taylor Series is also defined by the number of terms used in the Taylor Series, that is, the maximum order of derivatives in the Taylor Series. The Taylor series of a function $f(x)$ at any number k is:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^n(k)}{n!}(x-k)^n \qquad (2)$$

$f^n(k)$ symbolizes the nth derivative of f at the point k while $n!$ is representing the factorial of n. Taylor Series of order 0 is the function itself as the zero-order derivative of f is f and the terms 0! and $(x-k)^0$ are both evaluated to be 1.

Let us return to our problem just after the brief overview of Taylor Series. We will use Taylor Series Expansion. Left hand side function of the equation is expanded by using the Taylor Series Expansion - HOT represents higher order terms - [19]:

$$I(x+\delta x, y+\delta y, t+\delta t) = I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t + HOT \qquad (3)$$

HOT is cancelled by using first order Taylor Series. Then, Taylor Series Expansion found at the right hand side of the second equation is replaced in the LHS of the first equation:

$$I(x,y,t) + \frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = I(x,y,t) \tag{4}$$

The cancellation of $I(x,y,t)$ in both sides and dividing all sides by δt results in:

$$\frac{\partial I}{\partial x}\delta x + \frac{\partial I}{\partial y}\delta y + \frac{\partial I}{\partial t}\delta t = 0 \tag{5}$$

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t}\frac{\delta t}{\delta t} = 0 \tag{6}$$

$$\frac{\partial I}{\partial x}\frac{\delta x}{\delta t} + \frac{\partial I}{\partial y}\frac{\delta y}{\delta t} + \frac{\partial I}{\partial t} = 0 \tag{7}$$

$\left\{\frac{\delta x}{\delta t}, \frac{\delta y}{\delta t}\right\}$ is the set of pixel changes of the point over time in two dimensions. Therefore, the elements of the set construct the optical flow vector $\vec{V}(V_x, V_y)$. $\left\{\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y}, \frac{\partial I}{\partial t}\right\}$ is the set of derivatives of image according to three dimensions (2D point, time). The set is represented by $\{I_x, I_y, I_t\}$. The last equation is rewritten as:

$$I_x V_x + I_y V_y + I_t = 0 \tag{8}$$

$$I_x V_x + I_y V_y = -I_t \tag{9}$$

$$\nabla I . \vec{V} = -I_t \tag{10}$$

Now, the problem converges to solution of $\vec{V}$. The solution will be the estimation of optical flow. As there are two unknowns in the equation, it cannot be solved;

additional constraints and approaches are needed for solution. This problem is known as *aperture problem*. The problem is shown in Figure 3.4 with the motion direction.

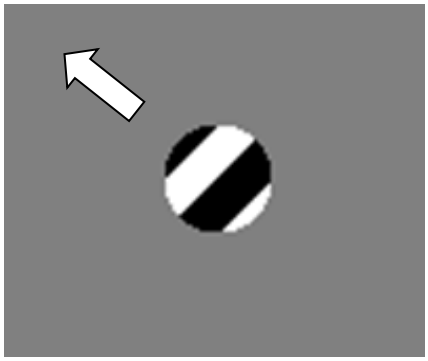


Figure 3.4: Aperture problem - 1

In the figure, there is a motion perception of a drawing beyond the grey curtain towards the top left corner. But, both the shape and motion direction of the drawing are not precise as we look at only a small part of the drawing behind the curtain. There are three options of drawing and motion direction which gives the effect shown in Figure 3.4. The options are shown in Figure 3.5.

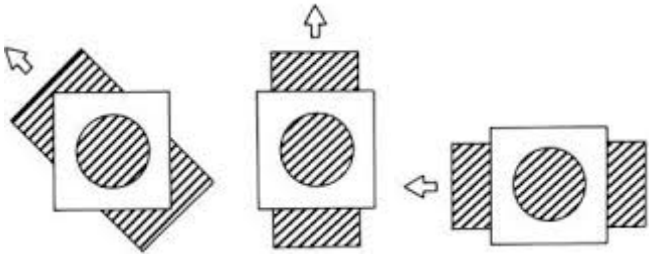


Figure 3.5: Aperture problem - 2

The solution of the aperture problem encountered in the equation will solve this dilemma and unify the observed motion options above.

3.1.2 Optical Flow Estimation Algorithms

Many algorithms using different approaches have been proposed for optical flow estimation. According to [17], optical flow estimation algorithms can be grouped according to the theoretical approach while interpreting optical flow; *differential techniques, region-based matching, energy-based methods and phase-based techniques*.

3.1.2.1 Differential Techniques

Differential techniques utilize velocity estimation from spatial and temporal derivatives of image intensity [17]. They are based on the theoretical approach proposed by [19]. The proposed approach results in the equation as it is mentioned above: $\nabla I.\vec{V} = -I_t$. Differential techniques try to develop methods for solving the problem generally represented by this equation.

Horn-Schunck method [19] is a fundamental method among differential techniques as the theoretical background of differential techniques is constructed by Horn-Schunck approach. Global smoothness concept is also integrated to the approach. Lucas-Kanade method [24] is also an essential method solving the mentioned differential equation for a set of neighboring pixels together by using a weighted window. [25 and 29] use second order derivatives generating the optical flow equations. Global smoothness concept is also used as well as the Horn-Schunck method. [18] proposes a distance based method efficient for real-time systems. The method is analyzed according to time-space complexity and its tradeoff. [20] suggests a classical differential approach. But, it is combined with correlation based motion descriptors. Then, the method tries to solve the discontinuities in optical flow.

3.1.2.2 Region-Based Matching

Region-based matching approaches alternate the differential techniques in case differentiation and numerical operations is not useful due to noise or small number of

frames [17]. In region-based matching, the concepts such as velocity, similarity, etc. are defined between image regions.

[21 and 27] propose region-based matching methods for optical flow estimation. In [21], the matching is based on Laplacian pyramid while [27] recommends a method based on sum of squared distance computation.

3.1.2.3 Energy-Based Methods

Energy-based methods are based on the output energy of filters tuned by the velocity [17]. [26] proposes an energy-based method fitting spatiotemporal energy to a plane in frequency space. Gabor filtering is used in the energy calculations.

3.1.2.4 Phase-Based Techniques

Different from energy-based methods velocity is defined as filter outputs having phase behavior. [22, 23 and 28] are the examples of phase-based techniques using spatiotemporal filters.

3.1.3 Feature Selection for Optical Flow Estimation

As it is mentioned before, selection of features is essential for the performance of optical flow estimation. It is important to select descriptive features. From this point of view, good features are the features that can be tracked clearly. Namely, the features should be invariant to noise, scale or illumination. There are many algorithms aiming to extract “good” features from images. Corners, high contrast region features like edges and relatively constant points (such as object corners, as their relative positions are unchanged despite motion) are some common nature of the “good” features.

Corner detectors are widely used in optical flow field. [30] is one of the earliest corner detection algorithms. Moreover, this study is the founder of “interest point”

concept. In this method, a “cornerness” measure is proposed. For each pixel, the values for cornerness are evaluated and the cornerness decision made according to a threshold. [31] offers a corner detector using a local auto-correlation function which measures the local changes of a pixel in different directions. According to the eigen values obtained from the auto-correlation matrix, corners are decided. [32] proposes a method based on [31]. While [31] uses simple translation, [32], differently, uses affine transform for modeling the motion. The method is well known as Shi-Tomasi algorithm.

Another approach for feature extraction is to find specific features as interest points. [33] proposes SIFT (Scale Invariant Feature Transform) algorithm to describe local features of an image. In this algorithm, the aim is to find detectable features which are invariant to scale and noise. Key locations are selected according to a Gaussian function applied in scale-space. These key locations are used in candidate object matching.

3.2 Temporal Representation Proposed Based on Optical Flow

In this thesis, an optical flow based temporal video information representation is proposed. This representation schema is actually a generic construction of continuous video parts with optical flow features and will be utilized for both segment representation in temporal video segment classification and cut representation in temporal video segmentation. Optical flow vectors are needed to be calculated for the selected sequential frames.

Optical flow estimation is important as the basic element of the model is optical flow vectors. As it is mentioned in part 3.1, detection of features and estimation of optical flow according to these features are the steps of optical flow estimation. The methods and approaches for both steps were discussed.

In our approach, *Shi-Tomasi* algorithm proposed in [32] is selected for feature detection. As it is mentioned before, Shi-Tomasi algorithm is based on Harris corner detector [31] and finds corners as interest points. Then, first, the Harris corner detector is shown:

$$S(u,v) = \sum_x \sum_y w(x,y)\,[I(x+u, y+v) - I(x,y)]^2 \tag{11}$$

The operator works by looking to the points through a window. S is the function representing the sum of squared difference between two windows of the operator. The function depends on the variables (u, v) representing the window movement in $x - y$ directions. w *i*s the weighting function while I is the intensity function. The purpose is to minimize S, that is, sum of squared difference. Using first order Taylor Series Expansion of the above equation:

$$S(u,v) = \sum_x \sum_y w(x,y)\left[I(x,y) + uI_x + vI_y - I(x,y)\right]^2 \tag{12}$$

$$S(u,v) = \sum_x \sum_y w(x,y)\left[uI_x + vI_y\right]^2 \tag{13}$$

Matrix representation of the above formula is:

$$S(u,v) = \sum_x \sum_y [u\ v] w(x,y) \otimes \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} \tag{14}$$

The Harris matrix [31] is obtained as follows:

$$M = \begin{bmatrix} I_x^2 & I_xI_y \\ I_xI_y & I_y^2 \end{bmatrix} \tag{15}$$

Shi-Tomasi algorithm uses the eigenvalues of the Harris matrix. In this context, it differs from Harris corner detector. λ_1, λ_2 are the eigenvalues of M. The cornerness measure is as follows:

$$\mathrm{C} = \min(\lambda_1, \lambda_2) \tag{16}$$

The algorithm assumes that minimum of two eigenvalues of Harris matrix determines the cornerness (C) of the point. Therefore, the corner decision is done using the eigenvalues of the matrix. Shi-Tomasi algorithm gives more accurate results compared with Harris detector. The algorithm is also more stable for tracking.

For estimating optical flow, Lucas-Kanade algorithm is selected [24]. With videos having sufficient information and excluding noise, Lucas-Kanade algorithm is successful. The algorithm works for the corners obtained from Shi-Tomasi algorithm. Basically, the following function should be minimized for each detected corner point as seen in differential approaches:

$$\in(\delta x, \delta y) = I(x, y) - I(x + \delta x, y + \delta y) \tag{17}$$

With suitable δx and δy optical flow vectors can be obtained. But, aperture problem is not solved yet with this minimization.

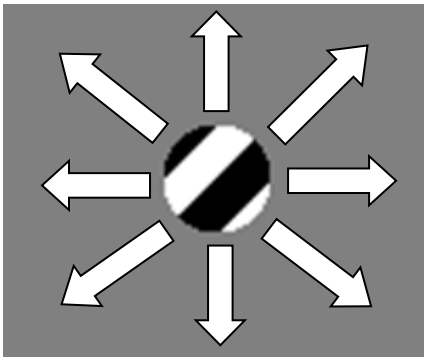


Figure 3.6: Solution for aperture problem

In Figure 3.4, aperture problem is explained practically. According to the explanation, in order to cope with the problem, the hole on the curtain which shows the background shape should be enlarged in all directions, like in Figure 3.6.

This practical approach should be reflected to the function definition $\in(\delta x, \delta y)$. The following definition updated with this approach solves the aperture problem:

$$\in(\delta x, \delta y) = \sum_{u-w}^{u+w} [I(x, y) - I(x + \delta x, y + \delta y)] \tag{18}$$

$$\in(\delta x, \delta y) = \sum_{u_y-w_y}^{u_y+w_y} \sum_{u_x-w_x}^{u_x+w_x} [I(x, y) - I(x + \delta x, y + \delta y)] \tag{19}$$

Summation on $x - y$ direction is a solution for the aperture problem. By using a window w centering the point (x, y), the estimation of optical flow of the point is extended with the neighboring points.

In our approach, Lucas-Kanade algorithm is applied to the corner points detected with Shi-Tomasi algorithm. The optical flow estimation will be handled by the Lucas-Kanade and Shi-Tomasi algorithm pair in the rest of the thesis. The selection is based on the facts about these algorithms mentioned in this chapter.

Video frames are selected according to a frequency of 6 frames/sec - 30 fps videos are used - from "*Hollywood Human Actions"* dataset [9]. In Figure 3.7, two frequently sequential frames obtained from the mentioned dataset are shown. There are more sequential frame examples from the dataset in Appendix B.

Figure 3.7: Consecutive frames for optical flow estimation

Figure 3.8 shows the optical flow vectors estimated for the detected points in the former video frame in the sequence. In Appendix B, more examples of optical flow estimations represented on different frames are provided.

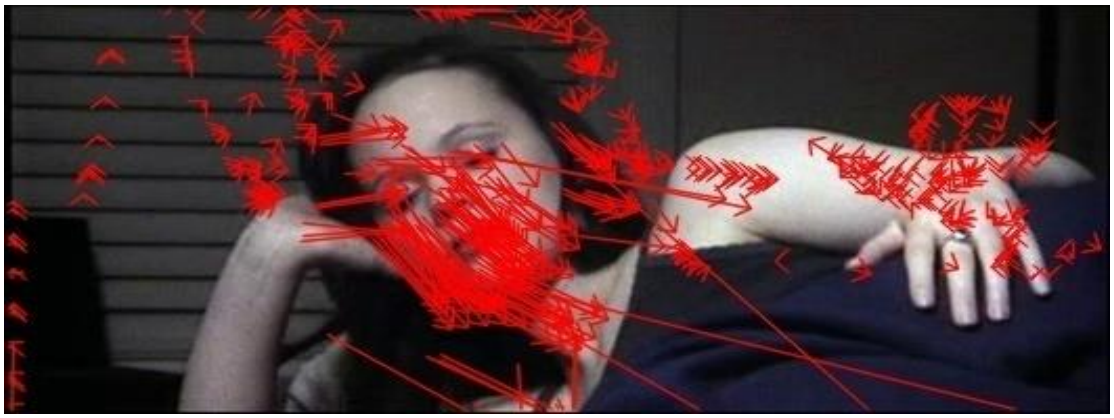


Figure 3.8: Frame with optical flow vectors

In our method, optical flow vectors are calculated for every detected point in all frequently selected frames. The set of optical flow vectors is the temporal information source for our representation.

The model below forms the back bone for our representation formalism. Optical flow vector set with an operator constructs the representation.

$$R = [S(V), \Phi] \tag{20}$$

$S(V)$ is the set of optical flow vectors while Φ is the descriptor operator. Operator defines the relation of the elements of the optical flow vector set of the frames. This relation exposes the temporal representation of video information. The operator may change according to the complexity of the model. It may vary from just counting the vectors to complex relations between the optical flow vectors or a combination of many relations. But, in basic, it contains optical flow vectors, optical flow histograms and optical flow sums. This generic representation can easily be adapted to different problems such as segment classification and cut detection. Choice of the operator and the optical flow representation may change drastically in different problems.

CHAPTER 4

TEMPORAL VIDEO SEGMENT CLASSIFICATION

Temporal video segmentation classification is a dimension of video action detection on which the video segments are classified according to predefined actions. A representation based on optical flow is proposed. The classification is done with SVM using that representation.

4.1 Proposed Model in Temporal Video Segment Classification

Usage of optical flow in video information representation is encountered in many studies including [35, 36, 37 and 60]. These studies propose novel methods motivating us for an optical flow based representation. Optical flow histogram is the most common way of optical flow based video representation. In [36], optical flow histograms are used for characterizing the motion of a soccer player in a soccer video. A motion descriptor based on optical flow is proposed and a similarity measure for this descriptor is described. The study of [37] uses optical flow by splitting it into horizontal and vertical channels. The histogram is calculated on these channels. Each channel is integrated over the angularly divided bins of optical flow vectors. In [35], histogram of oriented optical flow (HOOF) is simply used according to angular segments for each frame. The feature vectors are constructed with these angular values and combined for all frames of the video segment. The essential part for contribution here is the classification method. The classification is done with a proposed novel time-series classification method including a metric for comparing optical flow histograms. [60] proposes abnormal event detection algorithm. The

algorithm is based on Histograms of the Orientation of Optical Flow (HOOF) as the motion descriptor for a one class SVM classifier.

In our approach, histogram based optical flow approaches are enriched with a newly defined velocity concept; *Weighted Frame Velocity*. The idea, here, is originated from the inadequacy of optical flow histograms for interpreting information. Using optical flow histogram is discarded as the most important drawback of using histograms in segment representation is that the histogram similarity does not always mean the real similarity for motion characterization. Optical flow vectors are divided into angular groups and according to these groups, optical flow vectors are summed and integrated with the new velocity concept instead of a histogram based approach.

Estimating the optical flow vectors for each frame is the first step. Then, the equation $R = [S(V), \Phi]$ giving the generic representation is handled and adapted to segment representation. In this aspect, Φ is the operator defining the relations between the optical flow vectors and giving their meaning for representing the video segment composed of the set of optical flow vectors $S(V)$.

In our adaptation of the above representation to segment representation, the description of Φ is essential. The parameters used in the model is shown in Table 4.1.

Table 4.1: Segment representation model parameters

Parameter	**Definition**
F	Set of frames in the video segment
$S(V_f)$	Set of optical flow vectors in frame f
$S(V_f, \alpha, \beta)$	Set of optical flow vectors having angle between $\alpha - \beta$ in frame f
$A(\alpha, \beta)$	Weighted frame velocity of the whole segment direction having angle between $\alpha - \beta$
$\tau_f(\alpha, \beta)$	Threshold function for optical flow vectors having angle between $\alpha - \beta$ in frame f
$V(r, \angle\varphi)$	Optical flow vector having magnitude r and angle φ

The parameters above are the basic building blocks for constructing the representation model and the descriptor operator Φ. The following definitions are done for this purpose.

The definition of $S(V_f, \alpha, \beta)$ is made as follows:

$$S(V_f, \alpha, \beta) = \{V(r, \angle\varphi) \in V_f \mid \alpha < \varphi \leq \beta\} \quad (21)$$

Let's assume that $|F| = n$, m is the number of angle intervals and l is the length of the video segment in terms of seconds. With these assumptions, the representation of a video segment using average of optical flow vectors with angular grouping can be formulized as:

$$R = [\, \| \sum_{S(V_{\forall f}, \alpha_1, \alpha_2)} V(r, \angle\varphi) \|, \| \sum_{S(V_{\forall f}, \alpha_2, \alpha_3)} V(r, \angle\varphi) \|, \ldots, \| \sum_{S(V_{\forall f}, \alpha_m, \alpha_{m+1})} V(r, \angle\varphi) \| \,] \quad (22)$$

Above representation is a vector representation composed m of dimensions each of which is the magnitude of the sum of optical flow vectors for an angle interval. This is the common way of optical flow representation except the usage of vectors instead of histograms. This representation is descriptive as it utilizes the movement of a segment in different angel intervals by using the vector sum and magnitude calculation. But, it lacks the temporal information in terms of velocity. This means that the flow details throughout the frame sequence are discarded by only looking at the resulting direction and magnitude information. If this vector is extracted for each frame and combined for solving the problem as it is done in [35], curse of dimensionality problem arises. The dimension of the resulting vector will be $mxnxl$. For a 30-fps video of 5 seconds length with 30 angular intervals, for example, a vector of 4500 features is obtained for representing a segment. Using a frequency filter of 0.2 - 6 frames selected from a second of the video - will decrease the

dimension into 900, but the problem will not be able to be solved yet. This yields to the need for tackling the curse of dimensionality problem as it is handled in [35] with the newly proposed time series classification method including the new distance metric for the feature vectors.

In our approach, we deal with enriching the representation with new information to make the temporal information more descriptive without causing the curse of dimensionality problem. For this purpose, a new component is needed for the above feature representation based on movement magnitude of the segment in different directions. Velocity is selected as the fundamental idea for the new component as the velocity of the frames strongly affects the nature of video motion such as in walk and run events. For this purpose, weighted frame velocity concept is put forward. Abstractly, the feature vector would be composed of distance-velocity pair by adding the velocity component.

Weighted frame velocity is a metric which measures the velocity of a segment in a given dimension. It is weighted with the vector count in its direction. Theoretically, weighted frame velocity is formulated inspiring from the general velocity calculation $V = \frac{\Delta d}{\Delta t}$:

$$A(\alpha,\beta) = \frac{\sum_{i=0}^{n-1}[\| \sum_{S\left(V_{f_i},\alpha,\beta\right)} V(r,\angle\varphi) \| \, . \, |S\left(V_{f_i},\alpha,\beta\right)|]}{\sum_{i=0}^{n-1} | S\left(V_{f_i},\alpha,\beta\right)|} \tag{23}$$

The equation above calculates the weighted distance for each angular interval of each frame. Weight concept, here, is the weight of the frame to the segment. The weighted distances are summed up and averaged according to the number of vectors in the segment. The resulting value is the weighted velocity of the frames.

This approach, has another problem. As the velocity is weighted according to the number of vectors in the given angle interval of the frame, the noise or errors resulting from optical flow estimation and insignificant number of vectors in one

dimension unfairly dominate the values of that feature. In order to avoid this problem, thresholding is used as a common way of noise reduction. Therefore, a threshold function depending on the frame and angle interval is proposed to be used in the weighted frame velocity function.

$$\tau_{f_i}(\alpha,\beta) = \begin{cases} \dfrac{S\left(V_{f_i},\alpha,\beta\right)}{S\left(V_{f_i},0,2\pi\right)}, & \dfrac{S\left(V_{f_i},\alpha,\beta\right)}{S\left(V_{f_i},0,2\pi\right)} < C \\ 1, & otherwise \end{cases} \tag{24}$$

The above function is based on the ratio of the optical flow vectors of the given angle interval for the given frame. This ratio's being smaller or bigger according to the threshold value C directly determines the result of the function. At this point, estimation of threshold becomes important. The estimation will be done during the classification phase.

The updated weighted frame velocity function is updated with the thresholding function.

$$A(\alpha,\beta) = \frac{\sum_{i=0}^{n-1}[\,\| \sum_{S\left(V_{f_i},\alpha,\beta\right)} V(r,\angle\varphi)\,\| \,.\, \left|S\left(V_{f_i},\alpha,\beta\right)\right| \,.\, \tau_{f_i}(\alpha,\beta)]}{\sum_{i=0}^{n-1} |\, S\left(V_{f_i},\alpha,\beta\right)|} \tag{25}$$

The function affects the weighted contribution of each frame into the velocity of the segment in an angle interval according to its vectors' being noise or not.

As it is mentioned before, weighted frame velocity will be a new component to the feature vector representation based on the movement of the segment. Thus, the new representation is as follows:

$$R = [A(\alpha_1,\alpha_2), \| \sum_{S(V_{\forall f},\alpha_1,\alpha_2)} V(r,\angle\varphi)\,\|, A(\alpha_2,\alpha_3), \| \sum_{S(V_{\forall f},\alpha_2,\alpha_3)} V(r,\angle\varphi)\,\|, \dots, A(\alpha_m,\alpha_{m+1}), \| \sum_{S(V_{\forall f},\alpha_m,\alpha_{m+1})} V(r,\angle\varphi)\,\|\,]$$

(26)

Now, the operator Φ in the generic optical based representation model $R = [S(V), \Phi]$ is defined in this specific problem. The operator maps the optical flow vector set $S(V)$ to the feature vector R for a video scene.

$$\Phi: S(V) \rightarrow R \tag{27}$$

The function of the above mapping is shown in the obtained final representation. Mainly, it constructs the representation by applying the operator to the optical flow vectors. The operator Φ, in fact, is the symbolic representation of our method.

The practical usage of the representation is classifying the segments. The representation is for each video segment and has the size $mx2$. Segment classification, constant estimations and experiments with results and comparisons will be held in Chapter 6.

CHAPTER 5

TEMPORAL VIDEO SEGMENTATION

Temporal video segmentation is the problem of splitting the video information temporally into coherent scenes. Temporal video segmentation is generally originated from the needs of video segment classification. In order to classify the video scenes semantically as segments, they are needed to be extracted from whole video information in many cases. On the other hand, in some cases, video segment classification and temporal video segmentation are held together. This kind of methods follows an integrated approach by trying to carry out the scene extraction with the classification of related semantic information.

As temporal video information is composed of visually complicated and continuous sequence of video frames, analyzing the temporal boundaries of video events, actions, etc. is an important field of study. From this point of view, event boundary detection, temporal video segmentation, cut detection etc. are considered as similar concepts dealing with this problem.

Textual and audio features together with visual features are important sources of information for temporal video segmentation as in temporal video segment representation. Our main concern about automaticity and dependency of textual features to manual creation is also relevant here. Therefore, visual features' domination proceeds in this problem, too.

Temporal video segmentation methods are needed to be classified, as the methods can handle this problem from different point of views. Because we are dealing with

visual feature based segmentation, the methods should be analyzed and grouped accordingly.

5.1 Video Segmentation Approaches

Concerning visual features, there are many approaches dealing with temporal video segmentation. These approaches can be grouped according to the visual features used and methods followed. The groups are [47] *pixel difference based, histogram comparison based, edge oriented, motion based, and statistical feature based*.

5.1.1 Pixel Difference Based Approaches

Pixel difference based methods use the pixel intensity or color differences between the frames in order to characterize the cuts between the video scenes. Despite some additional improvements, they are generally based on the sum of differences. These improvements may vary from spatial segmentation of the frames to statistical analysis of the (distribution etc.) pixel color differences. Thresholding is also another important instrument in these approaches. Comparison of pixel differences with threshold values is a common way of detection. Basically, the detection is based on the following simple formula [47]:

$$D_{t,t+1}(i,j) = \begin{cases} 1, & \sum_{i,j} |I_t(i,j) - I_{t+1}(i,j)| > T \\ 0, & otherwise \end{cases} \quad (28)$$

Above formula shows the idea of the pixel different based approaches. Naturally, it is not a generic solution or method. This idea is the underlying factor shaping the methods.

[8] is an example method using pixel-to-neighbor image differences in the videos. The cut detection is implemented according to a ratio criterion. A dissolve detection

method is presented. It is based on an image difference and linearity error in a sequential image set.

Despite its simplicity and time efficiency, there is an important drawback of pixel difference based approaches; They are sensitive to motion [47]. As the changes in the pixel values are very fast and distant in video segments having motion, the pixel intensity differences may mislead the method. Then, sharp motion flows may be interpreted as cut points. Briefly, color analyzes is not reliable in most of the conditions.

5.1.2 Histogram Comparison Based Approaches

The histogram comparison based approaches use the color histogram differences between the frames. They are based on the comparison of the differences according to a metric.

Most of the histogram based approaches use thresholding. Thresholding works on the detection of the cuts by comparing the histogram differences with an estimated threshold. The detection is based on the formula [47]:

$$H(D_t, D_{t+1}) = \begin{cases} 1, & \sum_{i \in I} |H_t(i) - H_{t+1}(i)| > T \\ 0, & otherwise \end{cases} \quad (29)$$

Above formula summarizes the general idea of histogram differences. For all intensity values, difference of histogram items is calculated. Then, a distance metric can be proposed for the histogram differences. Histogram intersection is also a metric for evaluating the difference. According to the metric, difference values are compared to the estimated threshold.

Histogram comparison based approaches are successful in the cases independent of motion. The most important drawback of this approach is caused by the meaning of

histogram. Histogram similarity, in many cases, does not mean the real similarity in the context. Especially, in motion context, color histogram is not as effective as we expected. While different contexts may have similar histograms according to the color distribution, similar but extremely changing continuously contexts may extremely differ in terms of histogram similarity.

5.1.3 Edge Oriented Approaches

Edge oriented approaches of temporal video segmentation deal with the measurement of edge related metrics in video frames. These methods use two major concepts.

- Edge change ratio
- Edge histogram

Edge change ratio is a common concept in temporal video segmentation. It is calculated between two frames. After the detection of edges in both frames, edge changes is evaluated. The calculation is done using the entering and exiting edge pixels in successive frames. The generic approach is based on the following formula [77]:

$$ECR_f = \max\left(\frac{E_f^{out}}{E_f}, \frac{E_{f+1}^{in}}{E_{f+1}}\right) \tag{30}$$

This formula compares the ratio of exiting edge pixels in the former frame with the ratio of entering edge pixels in the letter frame. The maximum of two is selected as ECR of frame f. This ratio is then used to characterize the cut frames in temporal video segmentation.

Edge histogram is the histogram of the edge pixels. It can be used in the same way with color histograms. The change on edge histograms is calculated and the segmentation decision is done according to these histogram differences.

The drawback of edge oriented approaches is being sensitive to high speed motion [47]. As the edge ratios or histograms are changing drastically in high speed motion, classification of edge based changes become difficult.

5.1.4 Motion Based Approaches

Motion based approaches use the motion features in video frames. These approaches are based on the fact that the motion breaks mean cuts that can be used for motion detection and tracking.

Motion features varies from simple corners to 3D (with time dimension) interest points and optical flow. Their descriptiveness changes according to the motion information they are carrying. While the corner feature carries the stable point information in terms of intensity, 3D interest points behaves as descriptors defining the 2D points tracking over time throughout the consecutive frames.

[40 and 41] are important examples for motion based approaches. Camera motion analysis and optical flow estimations are used in these studies for cut detection.

The most important drawback of these approaches is the computational cost [47] and success rates on the extraction and tracking of motion features. In order to improve the success rates, many studies proposing novel methods are carried out about defining new motion descriptors. Clearly, the purpose is to store more motion information to the features for increasing the descriptiveness.

5.1.5 Statistical Feature Based Approaches

Statistical feature based approaches use the features (such as variance, standard deviation, and so on) of the visual features in video frames. In most of the methods, video frames are divided spatially into blocks and statistical parameters are calculated for each of these spatial locations. The spatial locations construct the feature vector for the related frame. Similarity metrics are proposed according to the feature vectors. The approaches are generally focused on motion detection and tracking using the motion features.

Statistical feature based approaches are similar to motion based approaches, in terms of computational cost and success rates. Dealing with success rates, new statistical models and motion features are proposed in statistical feature based methods.

In our study, we propose a method that allows us to enrich a pixel-based approach with a novel motion-based one. Optical flow is the motion feature used for describing a motion concept which integrates the motion information to a pixel difference-based metric.

5.2 Proposed Model for Temporal Video Segmentation

Temporal video segmentation is a direct field of interest in this thesis as video scenes for segment classification are needed to be extracted from the whole video. In this aspect, segments are obtained by detecting the cuts between the scenes.

Our problem is reduced to cut detection problem. Cut detection is a commonly studied field for video information retrieval purposes. Optical flow is an essential motion feature used in cut detection as well as in temporal segment representation. Key studies, [39 and 41], propose cut detection methods based on optical flow. The method utilizes linear prediction in order to detect cuts in different transition conditions. Each frame is spatially separated into 4x4 independent blocks. This

method is a kind of search model enabling us to detect the cuts using backward block searching for a match. Three sequential frames with the same matched block is taken into account where the fourth frame constitutes a subsequence. Mismatched blocks represent a change in block locations. The method has the ability for detecting cuts in many kinds of environment including quick scene changes, static scenes, and scenes with slow motion. [41] proposes a novel method for temporal segmentation of a video into scenes based on motion features. Markov Random Field is constructed for ensuring temporal continuity. Optical flow is the key motion feature used in the method. The approach is based on the fact that, optical flow changes differ at shot boundary points. Using the optical flow magnitudes, the aim is to detect the outliers.

Optical flow is the key concept behaving as an operator used in the pixel difference calculations in our method. The fundamental idea, here, is that some sort of change in optical flow character determines the cuts. We claim that the difference of intensity values between the pixels - mapped with optical flow vectors - of consecutive frames changes at the cut points. Calculated optical flow vectors in the first phase, video segment representation, are also used here as building block features operating on pixel difference calculations to represent scene changes. This yields to a decrease in the computational complexity because of the fact that the feature base, optical flow vectors, is same and singularly calculated for both phases.

The proposed method partially resides in the group of pixel difference based approaches. This is the fundamental idea in our approach. Our method combines the pixel based cut detection methods with the motion based ones in order to get rid of the main drawback (being sensitive to motion) of pixel based methods. Practically, the motion base approach is proposed by integrating to the pixel difference based approach. Accordingly, *average motion vectors* are calculated based on optical flow vectors and also pixel matching based on these average vectors are constructed for consecutive frames. Video frame transition is modeled based on this pixel matching. Accordingly, all frame transitions form the cut candidate set. Therefore, binary cut classification can work on this constructed set. In order to have a more descriptive

feature vector, the frames are spatially divided into blocks and the mentioned calculations are done based on blocks instead of frames. The method is also improved by other additional updates.

Estimated optical flow vectors for each frame in the first step are used. Eq. (20) giving the optical flow based generic representation, defined in Chapter 3, is handled and adapted to cut detection. From this point of view, Φ in Eq. (20), is the operator defining the relations between the optical flow vectors $S(V)$ and giving their meaning for representing cuts.

In our adapted method, the temporal representation of model parameters is defined as in Table 5.1.

Table 5.1: Temporal representation model parameters

Parameter	**Definition**
$P_{m,i}$	Spatial block m in i^{th} frame
$S(V_{P_{m,i}})$	Set of optical flow vectors in block $P_{m,i}$
$V'_{P_{m,i}}$	Average motion vector in block $P_{m,i}$
$D_{P_{m,i}}$	Block distance of block $P_{m,i}$ with $P_{m,i+1}$
$px_j^{P_{m,i}}$	j^{th} pixel in block $P_{m,i}$
$O(px_j^{P_{m,i}})$	Location of optical flow pixel $px_j^{P_{m,i}}$ in $(i+1)^{th}$ frame
$\mid S \mid$	Cardinality of set S
$I(px)$	Intensity value of pixel px

The first step in our method is to partition the frames into spatial locations. These locations will hold the visual features. By grouping the optical flow vectors according to the spatial locations, the information carried by the spatial locations becomes color information with optical flow vectors. In detail, this grouping means that the pixels holding optical flow vectors are grouped.

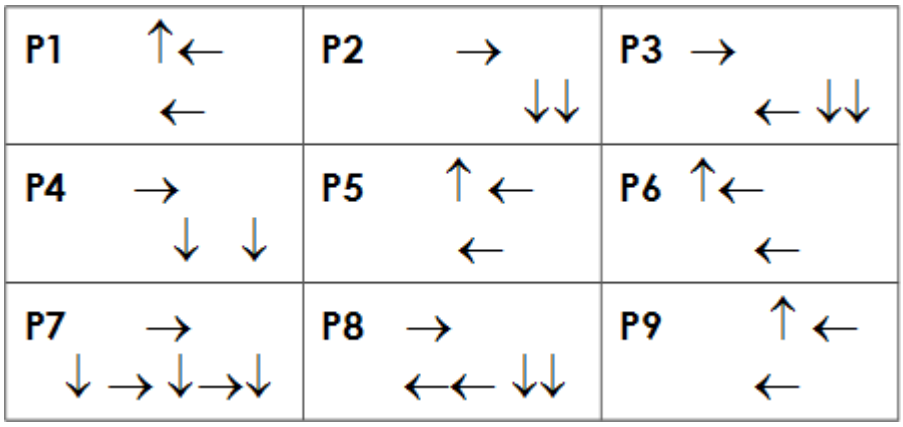


Figure 5.1: Spatial partitioning of video frame

Figure 5.1 shows a partitioning of size $3x3$. The optical flow vectors are shown inside the locations.

The parameters defined in Table 5.1, are the basic building blocks for constructing the model and the descriptor operator Φ. The definitions and formulas of the method utilize them.

The definition of $V'_{P_{m,i}}$ is as follows:

$$V'_{P_{m,i}} = \frac{\sum_{v \in S(V_{P_{m,i}})} V(r, \angle\varphi)}{\left|S(V_{P_{m,i}})\right|} C \tag{31}$$

Above representation, describing average motion vector concept, calculates the sum of optical vectors in a block of the related frame. The sum is averaged by the total number of optical flow vectors in that partition. C is a threshold between $0-1$ which is used for avoiding noise. It depends on but is not equal to the ratio $\frac{\left|S(V_{P_{m,i}})\right|}{|S(V)|}$.

Average motion vector is used in the pixel mapping part. A metric is needed for calculating the differences between the blocks of sequential frames. Euclidean distance metric is used to calculate the distance between frame blocks. Assuming that

each block of a video frame includes N pixels, the following formula defines the distance accordingly.

$$D_{P_{m,i}} = \sqrt{\sum_{j=0}^{N-1} \left| I\left(px_j^{P_{m,i}}\right) - I\left(px_{j'}^{P_{m,i+1}}\right) \right|^2} \tag{32}$$

This formula is based on the Euclidean distance between the intensity values of the mapping blocks of consecutive frames. Mapping block in the letter frame is constructed from the optical flow vectors of the related block in the former frame. The mapped pixel originated from pixel j of former frame block is represented as $px_{j'}^{P_{m,i+1}}$ in the above representation.

The mapping function for frame blocks is as follows:

$$px_{j'}^{P_{m,i+1}} = \begin{cases} O\left(px_j^{P_{m,i}}\right), & if\ px_j^{P_{m,i}}\ is\ a\ part\ of\ an\ optical\ flow\ vector \\ px_j^{P_{m,i}} + V'_{P_{m,i}}, & otherwise \end{cases} \tag{33}$$

According to the above function and distance formula, distance values are expected to differ in cut points. Then, the feature vector is represented by using the $D_{P_{m,i}}$ values. First, distance is calculated for each block in the frame and combined for the vector representation. Assuming that the frames are partitioned into M parts, frame transition can be represented as:

$$F_{i,i+1} = [D_{P_{1,i}}, D_{P_{2,i}}, D_{P_{3,i}}, \dots, D_{P_{M,i}}] \tag{34}$$

The vector set of the video is shown below assuming that the video has W frames:

$$S = \{F_{1,2}, F_{2,3}, \dots, F_{w-1,w}\} \tag{35}$$

Moreover, we follow a vector-based representation based on the similarity between the video fragments containing cuts and the sentences of natural language. Each word in a sentence has a general meaning similar to the video frames. But, the sentences composed of several words may have completely different meanings in many cases. Even if the case is not so, it is definite that the sentences contain more expanded meanings compared to the words composing the sentences.

In the case of video fragments containing cuts, the situation is comparable with the sentences regarding our base feature as optical flow. The cut frames may not have the precise meaning as the cut. But, they have a stronger meaning with their neighboring frames before and after in terms of optical flow nature. Herein, the sliding windows approach - widely used in text information extraction - is considered to take the neighboring frames into account. The approach and its implementation will be discussed with the results and contribution in Chapter 6.

Now, the operator Φ in the generic optical based representation model $R = [S(V), \Phi]$ which is proposed with Eq. (20) in Chapter 3 is redefined in this specific problem. The model is updated with the intensity values as follows:

$$R = [I, S(V), \Phi] \tag{36}$$

Φ operator, which contains the function defined in Eq. (10) maps pixel intensities I by the help of the optical flow vector set $S(V)$ to the feature vector R.

$$\Phi: \{I, S(V)\} \rightarrow R \tag{37}$$

The function of the above mapping is shown in the obtained final representation. Mainly, it constructs the representation by applying the operator to the intensities using optical flow vectors. The operator Φ, in fact, is the symbolic representation of our method.

The practical usage of the representation is classifying the cuts. The representation is for each consecutive frame pairs and has the size M. Cut detection, constant estimations and experiments with results and comparisons will be held in Chapter 6.

CHAPTER 6

EXPERIMENTS, RESULTS AND IMPROVEMENTS

6.1 Temporal Segment Classification

Temporal segment classification is carried out using the vector representation proposed in Chapter 4. Support Vector Machines (SVM) is used for non-linear classification. Gaussian radial basis function is used as SVM kernel.

$$K(x_i, x_j) = e^{\left(-\frac{1}{2\sigma^2}\|x_i - x_j\|^2\right)} \tag{38}$$

In Eq. (38), $K(x_i, x_j)$ is the definition of the kernel function for the feature vectors x_i and x_j. $\| x_i - x_j \|^2$ represents the squared Euclidean distance between the feature vectors x_i and x_j. σ represents the free parameter of the kernel function.

Hollywood Human Actions dataset [9], is used for evaluation. Hollywood dataset includes video segments composed of human actions from 32 movies. Each segment is labeled with one or more of 8 action classes: "*AnswerPhone, GetOutCar, HandShake, HugPerson, Kiss, SitDown, SitUp, StandUp"*. While, the test is set obtained from 20 movies, training set is obtained from 12 movies different from the test set. The training set contains 219 video segments with manually created labels. The test set contains 211 samples with manually created labels. The details of the data set are given in Table C.1, Appendix C.

After the optical flow estimations, the calculations for constructing feature vectors are carried out accordingly and feature vectors are obtained for the test data. The number of angular intervals is taken as 30 [35]. In order to estimate the threshold C in the threshold function discussed in Chapter 4, experiments are carried out. The threshold function is given below:

$$\tau_{f_i}(\alpha,\beta) = \begin{cases} \dfrac{S\left(V_{f_i},\alpha,\beta\right)}{S\left(V_{f_i},0{,}2\pi\right)}, & \dfrac{S\left(V_{f_i},\alpha,\beta\right)}{S\left(V_{f_i},0{,}2\pi\right)} < C \\ 1, & otherwise \end{cases} \tag{39}$$

The result of the experiments for determining the best threshold value is shown in Figure 6.1.

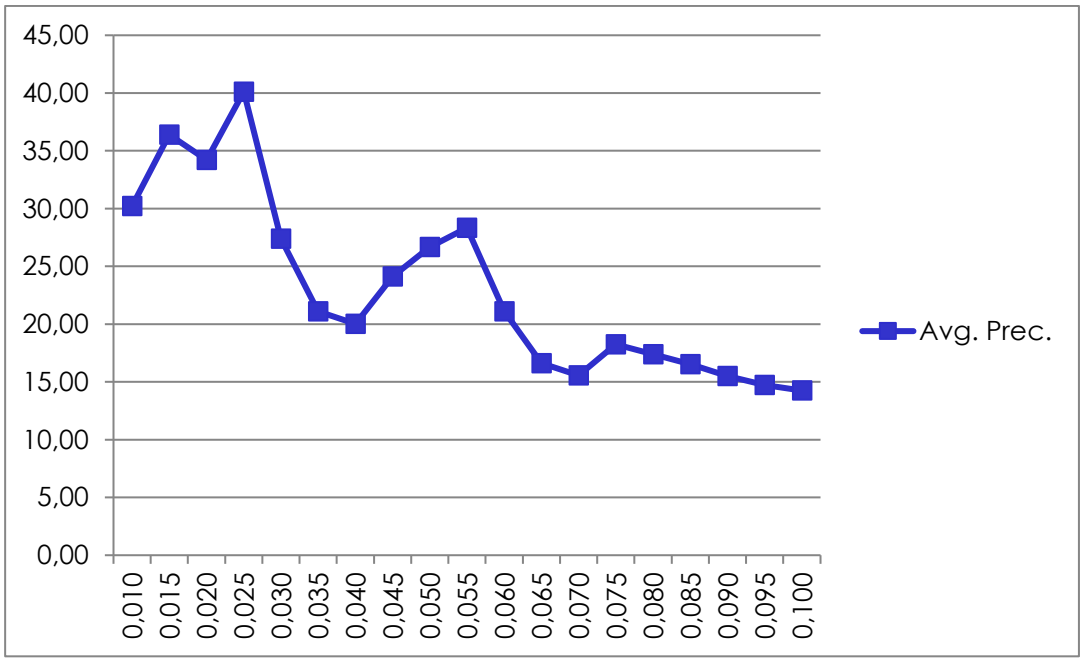


Figure 6.1: Threshold estimation in segment representation for Hollywood Human Actions

According to the figure, the threshold value 0.025 gives the best result in the experiment.

After, selecting the threshold, experimental results with this threshold value is detailed. The results are compared with the ones obtained from the study [9] in Table 6.1. This study proposes a method that can be assumed as a reference for video segment classification based on space-time points.

Table 6.1: Comparison of the results of video segment classification for Hollywood Human Actions

Action	**Recall**	**Precision**	
	Ours	**Laptev et al. [9]**	**Ours**
StandUp	79.2%	50.5%	40.0%
SitDown	88.9%	38.6%	42.9%
HandShake	94.7%	32.3%	40.0%
Hug	90.9%	40.6%	44.4%
SitUp	80.0 %	18.2%	33.3%

Recall and precision of the test is shown in the table. Recall values are high, as the ratio of each event is low in total. Except the standup action, better results are obtained in the precision comparison. Especially, concerning “SitUp” action, the success rate is doubled.

Another comparison is made with the popular state-of-the-art Weizmann data set. The data set contains the actions “walk”, “run”, “jump”, “side”, “bend”, “one-hand wave”, “two-hands wave”, “pjump”, “jack”, “skip”. The details of the data set are given in Table C.2, Appendix C.

First, the threshold estimation is carried out in this set again. The result is shown in Figure 6.2.

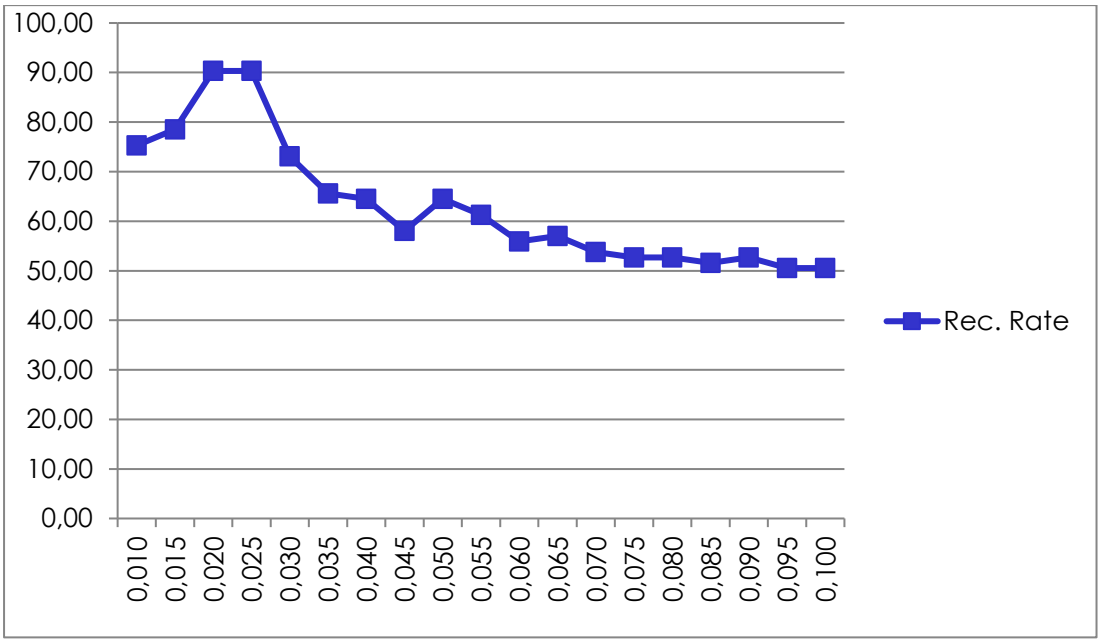


Figure 6.2: Threshold estimation in segment representation for Weizmann data set

According to the figure, the threshold values $0.020 - 0.025$ give the best results. These values are used in evaluating the results over this data set.

The comparison of our method in terms of recognition rates with the essential studies having different viewpoints in human action recognition and segment classification is shown in Table 6.2.

Table 6.2: Comparison of the results of video segment classification for Weizmann data set

Methods	**Recognition Rates**
Gorelick et al. [42]	97.83%
Chaudry et al. [35]	94.44%
Ali et al. [43]	92.60%
Ours	90.32%
Niebles et al. [44]	90.00%
Lertniphonphan et al. [15]	79.17 %
Niebles et al. [45]	72.80 %

The methods shown in Table 6.2 are some of the reference studies dealing with the temporal segment classification problem. [42 and 43] approaches propose methods using interest point features having time dimension. The classification is done according to these novel 3D features. [35 and 15] propose optical flow based classification. [35] focuses on representing the segments frame by frame optical flows with high dimensions causing curse of dimensionality problem. Instead of dealing with the representation, the method aims to contribute by finding new metrics and time series patterns in this high dimensional data. [15], on the other hand, proposes a general optical flow based representation structure based on direction histograms of optical flow field. [44 and 45] present models based on spatio-temporal words. Both methods see the segments as bag-of-features and makes the classification according to the code words based on interest points. While [44] learns the probability distributions of the spatio-temporal words by using Latent Semantic Analysis, [45] creates the bag-of-words model by selecting the bag-of-words features based on a hierarchical model.

When we analyze the results, methods proposing interest point based new 3D features are more successful than the other models. Methods focusing on generically mining the highly over-descriptive data in terms of time domain present high success rates as they develop the model independent from the contributions in video features. Our optical flow based method is more successful than the general optical flow based approach making contribution with optical flow based segment representation. It is also more successful than the bag-of-words based methods.

The details of the recognition results are shown in Table 6.3 according to the action type vs. method. Each value shows the recognition rate of the specified method calculated with its metric for the specified action. The last column shows the correctly classified action results of our method according to the set defined in Table C.2, Appendix C. The detailed results of [42] is excluded as the recognition rates of this study is calculated in terms of the space-cubes instead of action instances.

Table 6.3: Action based detailed results of temporal video segmentation for Weizmann data set

	Chaudry et al. [35]	Ali et al. [43]	Ours	Niebles et al. [44]	Lertniph. et al. [15]	Niebles et al. [45]	Actions (Ours)
Bend	1.00	1.00	1.00	1.00	0.67	1.00	9
Jack	1.00	1.00	1.00	1.00	1.00	1.00	9
Jump	0.89	0.56	0.89	0.78	1.00	0.78	8
Pjump	1.00	1.00	1.00	1.00	0.67	1.00	9
Run	0.89	0.89	0.90	0.89	1.00	0.56	9
Side	1.00	0.89	0.89	1.00	1.00	0.56	8
Skip	1.00	-	0.80	0.44	0.25	-	8
Walk	1.00	1.00	1.00	1.00	1.00	0.56	10
Wave1	0.67	1.00	0.78	0.89	1.00	0.44	7
Wave2	1.00	1.00	0.78	1.00	0.33	0.67	7

Generally, for all action types, high recognition rates are obtained in [42] defining actions as 3D space-time shapes. As it is assumed that the silhouette of a person at each time slice is determined precisely, the method is restricted and sensitive to specific cases. [35] also obtains better results in almost all cases. It is a more general method using optical flow and time-series analysis. As the approach proposes a time series analysis based method on the whole optical flow field, the advantage of using highly over-descriptive data allows the method to step forward. We have similar results with [43]. The difference is resulted from the exclusion of the skip action.

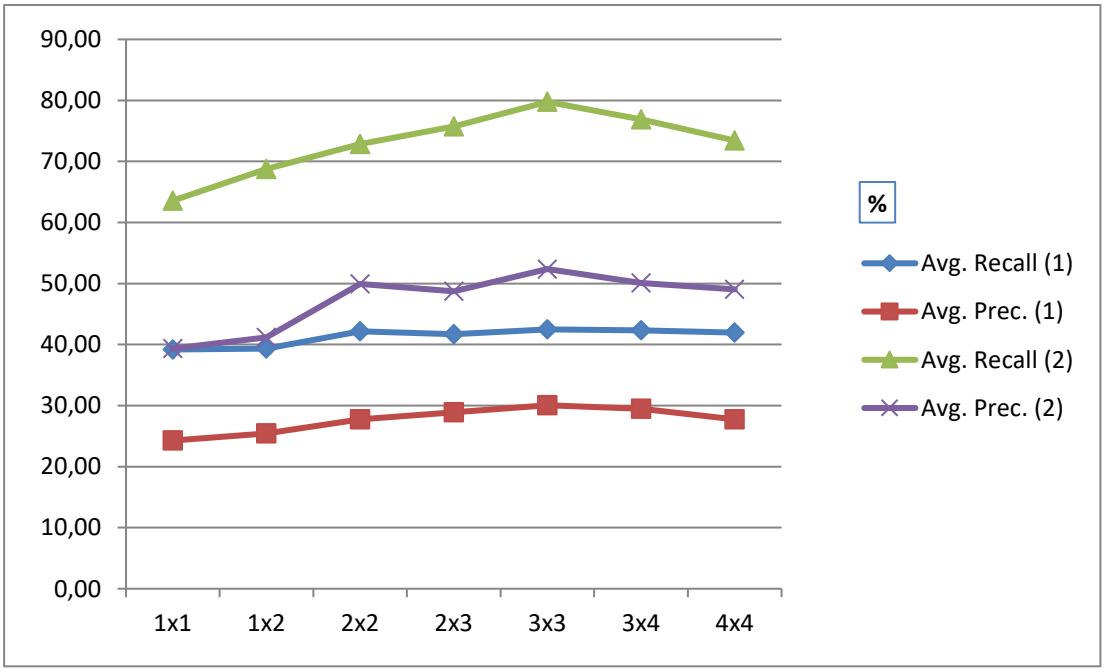
Our method is more successful than the general optical flow based approach [15]. The major difference is caused by skip action as their approach confuses it with run action. This is expected because the method may not discriminate these similar actions by using optical flow histogram. In this aspect, our method comes into prominence with its rich representation based on optical flow. Our method also performs better than the bag-of-words based methods [44 and 45] generally. The

major difference is the skip and wave actions especially compared to [45]. It is difficult to differentiate these similar actions using code words.

6.2 Temporal Video Segmentation

Temporal video segmentation is carried out using the vector representation proposed in Chapter 5. SVM is used for non-linear binary (cut/non-cut) classification as well as temporal segment classification. Gaussian radial basis function is also used as SVM kernel. Video Segmentation Project in Carleton University data set [48] is used for evaluation. The details of the data set are given in Table C.3, Appendix C. The set contains 10 different video data. The movies with motion and actions are selected. These are C, E, G, H, J. First, the size of spatial partitioning in the video frame is estimated. Here, the contribution of *average motion vector* concept is also tried to be shown by making a slight modification to the method. The modification is performed by removing the *average motion vector* concept from the related equation defined in Chapter 5.

Figure 6.3: Partition size estimation in temporal video segmentation for Video Segmentation Project in Carleton University data set



The results of the experiments for determining the best value for partitioning and showing the effect of average *motion vector* are shown in Figure 6.3. According to the figure, the optimal partition size is estimated as 3×3 regarding the cut detection recall/precision results. It is also seen that the results of the proposed method (2) are far better than the modified one (1). Therefore, the contribution of *average motion vector* is obvious. After selecting the partition size, detailed experimental results of the proposed method with this size is found and compared to the ones obtained from the study in [48], as in Table 6.4.

Table 6.4: Comparison of the results of temporal video segmentation for Video Segmentation Project in Carleton University data set

Set	**Recall**		**Precision**	
	Whitehead et al. [48]	**Ours**	**Whitehead et al. [48]**	**Ours**
C	87.0%	42.6%	59.5%	23.0%
E	100%	83.3%	93.8%	50.0%
G	94.4%	77.7%	81.0%	51.8%
H	89.5%	78.9%	89.5%	48.4%
J	89.7 %	79.3%	49.7%	57.5%

As we see in the table, both precision and recall rates are far from the ones in [48]. According to the analysis on the vectors, the separation could not be implemented under these conditions. In order to make the vector more descriptive sliding windows method is used. As the cut frame is a peak point compared to its neighboring frame transitions, neighboring frames should be used in the representation of the cut point.

Sliding windows methods are used widely in many areas. Especially, information extraction algorithms utilize this approach. The main idea is that the meaning of a word does not only depend on its meaning but also the meaning of its context, in other words the meanings of the words preceding and following it. An example is shown in Figure 6.4.

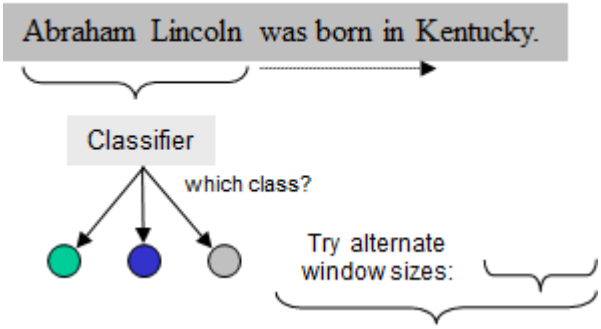


Figure 6.4: Sliding window method in IE

In Figure 6.4, meaning of a name is tried to be extracted. While the name itself does not give the desired meaning, dealing with chunks (neighboring words) by using sliding windows extracts the desired meaning.

Similarly, in cut frames, neighboring frames contribute to the meaning of the cut frame as mentioned before. The set of frame representation was proposed in Chapter 5:

$$S = \{F_{1,2}, F_{2,3}, \dots, F_{w-1,w}\} \tag{40}$$

By using a sliding window of size s, the new representation of a frame will be as follows:

$$F_i' = [\dots, F_{i-2,i-1}, F_{i-1,i}, F_{i,i+1}, F_{i,i+2}, \dots] \tag{41}$$

The size of the new vector is as follows:

$$|F_i'| = sx|F_{i,i+1}| \tag{42}$$

The frame set is constructed as follows:

$$S = \{F_1', F_2', \dots, F_w'\} \tag{43}$$

The experiments are carried out again with the new update to the method. First, the size of sliding window is estimated using the frame partition size $3x3$ estimated before. The same data set is used. The results are shown in Figure 6.5:

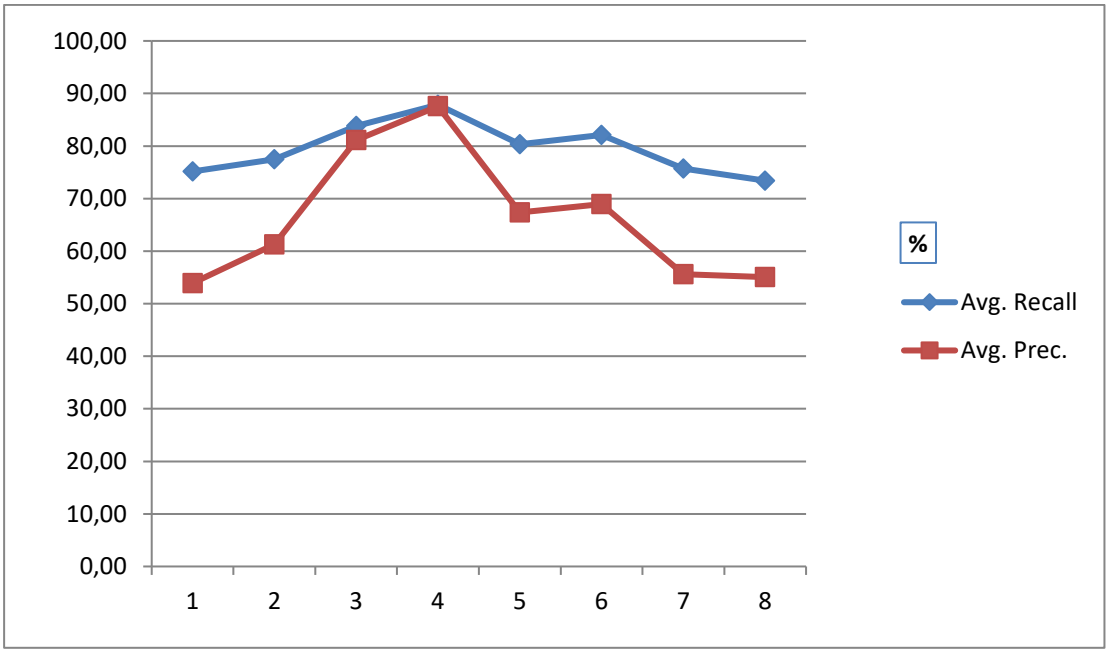


Figure 6.5: Estimation of sliding window size

According to the figure, optimal sliding window size is estimated as 4. After the experiments with the estimated window size, the results are seen and compared in Table 6.5.

Table 6.5: Comparison-1 of the results of temporal video segmentation using sliding window for Video Segmentation Project in Carleton University data set

Set	**Recall**		**Precision**	
	Whitehead et al. [48]	**Ours**	**Whitehead et al. [48]**	**Ours**
C	87.0%	57.4%	59.5%	36.0%
E	100%	100%	93.8%	96.6%
G	94.4%	94.4%	81.0%	94.4%
H	89.5%	92.0%	89.5%	90.2%
J	89.7 %	80.0%	49.7%	62.5%

Satisfactory results were obtained with these new improvements. [48] proposes a feature based cut detection with automatic threshold selection. This comparison shows that our approach based on optical flow representation is more successful in terms of recall and precision in most cases.

Our method is also compared with other successful state-of-the-art methods having different approaches with the same data set. The video having computer animations (J) is excluded to have a common context with the other methods. [50] uses histogram based edge tracking methods in cut detection while [49] proposes a fuzzy rule based approach tackling the cut detection problem. [41] introduces a dynamic threshold based method using optical flow for cut detection. The results of the comparisons are shown in Table 6.6.

Table 6.6: Comparison-2 of the results of temporal video segmentation using sliding window for Video Segmentation Project in Carleton University data set

Set	**Recall**				
	Whitehead et al. [48]	**MOCA [50]**	**Ours**	**Roghayeh et al. [49]**	**Kowdle et al. [41]**
E	100%	81.0%	100%	92.3%	93.3%
H	89.5%		92.0%	92.5%	94.7%
G	94.4%	61.1%	94.4%	88.9%	88.9%
Set	**Precision**				
	Whitehead et al. [48]	**MOCA [50]**	**Ours**	**Roghayeh et al. [49]**	**Kowdle et al. [41]**
E	93.8%	95.3%	96.6%	85.7%	94.2%
H	89.5%		90.2%	100%	88.8%
G	81.0%	91.7%	94.4%	100%	90.4%

In the table, it is seen that our method gives better results than the threshold oriented visual feature based, and edge histogram based methods [48 and 50]. It results from the fact that the motion effects are better handled by our method. The simple features used by thresholds and edge histograms cannot describe cut nature as good as our

motion based representation especially in the videos having motion. Moreover, the results are close to the ones in [49]. It is seen that, the rules defining the domain knowledge in the fuzzy rule based method affects the precision values positively. However, rule based approaches make the methods more domain specific. Last, we compare our method with another optical flow based method [41] and obtain better results in general. This approves the success of *average motion vector* concept with our descriptive representation formalism against the dynamic thresholding with average optical flow vector magnitude. The complexities are similar as both of the methods include optical flow calculations and window based analyses of optical flow vectors as the dominant cost operations. Our method is computationally more complex than the other methods except [41] because of the optical flow vector calculations.

CHAPTER 7

CONCLUSION AND FUTURE WORK

This study proposes frame segmentation and segment classification methods for action detection. The fundamental problem inspiring us is the representation of temporal information. In many fields, representation of temporal information is essential to retrieve the required information among a temporal set. The solution to the problem varies from representing the temporal entities in each time slice to represent a simple summary of whole time interval. Efforts for finding a solution between these two endpoints, require dealing with the problem from different point of views. This is because, the representation level, here, completely changes the problem. For instance, handling temporal information having high frequency over time, just like video frames in our study, requires representing all the information in all time slices, which in return, will cause the curse of dimensionality problem. Moreover, it will be very difficult to retrieve meaningful information from such redundant video data. There are studies concentrating on solving curse of dimensionality problem in temporal information. On the other hand, representing a single summary will cause the problem of lacking the flow of temporal information. In these cases, the focus of studies will be finding supportive information from different sources and integration of these sources in a singular representation.

This study aims to solve the temporal information representation problem in video domain. As the video data is a perfect example of high frequency temporal information, representation of video information is essential for the purposes emerging in information retrieval. Content-based video information retrieval is

selected as our specific domain. In this domain, various directions exist according to the video content concepts - object, action, event, etc. Action detection is our direction as the sub-domain. This problem domain is divided into two parts; temporal videos segment classification and temporal video segmentation.

The study is shaped on visual features of the video information for the automaticity concerns. As it is mentioned below, the representation level determines the reduced problem. In this context, our approach is to represent the video scenes by avoiding the lack of temporal information flow and without causing the curse of dimensionality problem. Therefore, using more descriptive and high level visual features having the ability to host the additional temporal nature of the simpler features such as color, edge, corner, etc. becomes unavoidable. This will inject the high load of temporal information residing in high dimensional representation to the mentioned high-level features.

Our decisions summarized above take us to the complex visual features having temporal dimension. In this research, we observed space-time related 3D features combining 2D features with temporal information. Space-time interest points and space-time shapes for actions are important examples of these features. We also researched high level features such as optical flow to describe the motion of frame features. Optical flow vectors are calculated and used in temporal video information representation. Curse of dimensionality problem may occur if most of the frames are represented using optical flow vectors. Thus, solving this problem also becomes necessary in such cases.

An optical flow based approach is proposed in this thesis for representing temporal video information by originating from the points above. The approach basically including optical flow based features is suggested. This generic approach is used in both temporal video segment classification and temporal video segmentation. The adaptation of the model to video segment classification with new concepts and relations is presented. The weighted frame velocity concept is put forward in order to

strengthen the representation with the velocity of video frames. This representation formalism is tested with SVM based classification of video segments. The results are promising.

Another part of action detection is temporal video segmentation. In this thesis, the problem is converted to cut detection. As we have the representation based on optical flow, the calculated optical flow vectors and their relations are utilized. The generic approach for representation is adapted with new concepts and relations for cut detection.

Based on the fact that optical flow changes differ at shot boundary points, an optical flow based cut detection model is proposed. Average motion vector concept is proposed for video frames. Using this concept and optical flow vectors, consecutive frame transitions are modeled in terms of differences. The model is tested using SVM based binary classifier having cut/non-cut decision. The results are compared with reference studies and contribution of the model is put forward with better success rates.

The main contribution of the methods is its solution for temporal video representation problem in action detection of content-based video information retrieval domain. In addition to the success rates, the methods have also less computational complexity considering the whole action detection process - segmentation and segment classification - as the calculated optical flows are used as the basic visual features in both phases.

As the future work, this thesis opens new paths in action detection. Considering our approach to action detection as temporal video segmentation and temporal video segment classification, it is important to complete the action detection process. Certainly, the study does not finish the whole action detection process. It takes up an important part, but challenges such as video action localization, detecting chaining actions and multi-modal approaches are the directions forwarding us. In temporal

video segment classification, localization of video actions is essential in order to cope with diverse real-world videos. Supervising data sources other than video data can be helpful in this part. Detecting multiple chaining actions is also important in real-world conditions. The problem of detection of chaining actions in a single scene can open up paths to different areas of research other than vision. Application of multi modal approaches is another challenge for future work. Additional data sources to video data and focusing on this supervision will also be useful for the whole detection process.

REFERENCES

[1] T. C. Vasileios, C. L. Aristidis and P. G. Nikolaos, "Scene Detection in Videos Using Shot Clustering and Sequence Alignment". *IEEE Transactions on Multimedia*, vol. 11, no. 1, 2009.

[2] A. Ghoshal, P. Ircing and S. Khudanpur, "Hidden Markov Models for Automatic Annotation and Content Based Retrieval of Images and Video". *Proc. of SIGIR*, 2005.

[3] L. W. Chang, W. N. Lie, and R. Chiang, "Automatic Annotation and Retrieval for Videos". *Proc. of PSIVT*, LNCS 4319, pp. 1030-1040, 2006.

[4] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra, "Video Event Classification Using String Kernels". *Multimedia Tools Application*, 48:69-87, 2009.

[5] F. Wang, Y. Jiang and C. Ngo, "Video Event Detection Using Motion Relativity and Visual Relatedness". *ACM Multimedia*, 2008,

[6] C. Ngo, T. Pong and H.Zhang, "Motion-Based Video Representation for Scene Change Detection". *International Journal of Computer Vision*, 50(2), 127-142, 2002

[7] P. Sand and S. Teller, "Particle Video: Long-Range Motion Estimation Using Point Trajectories". *International Journal of Computer Vision*, 2008.

[8] M. Luo, D. Menthon and D. Doermann. “Shot boundary detection using pixel-to-neighbor image differences in video”. *In TRECVID*, 2004.

[9] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, “Learning realistic human actions from movies”. *Proc. of CVPR*, Anchorage, US, 2008.

[10] A. Burton and J. Radford, "Thinking in Perspective: Critical Essays in the Study of Thought Processes". *Routledge*, 1978.

[11] D. H. Warren and E. R. Strelow, "Electronic Spatial Sensing for the Blind: Contributions from Perception". *NATO Science Series*.

[12] J. J. Ibson, "The Perception of the Visual World". *Houghton Mifflin*, 1950.

[13] C. S. Royden and K. D. Moore, "Use of speed cues in the detection of moving objects by moving observers". *Vision research*, 59, 17–24, 2012.

[14] C. Gianluigi and S. Raimondo, “An innovative algorithm for key frame extraction in video summarization*”. Journal of Real-Time Image Processing*, vol. 1, no. 1, pp. 69–88, 2006.

[15] K. Lertniphonphan, S. Aramvith, T. Chalidabhongse “Human Action Recognition using Direction Histograms of Optical Flow”. *In ISCIT*, 2011.

[16] J. J. Gibson, “The Perception of the Visual World”. Houghton Mifflin, 1950.

[17] J. Barron, D. Fleet and S. Beauchemin, "Performance of optical flow techniques". *International Journal of Computer Vision*, pp. 43-47, 1994.

[18] T. Camus, "Real-time quantized optical flow". *The Journal of Real-Time Imaging (special issue on Real-Time Motion Analysis)*, 3:71–86, 1997.

[19] B. K. P. Horn and B. G. Schunck, "Determining optical flow". *Artificial Intelligence*, no. 17, pp. 185-203, 1981.

[20] M. Proesmans, L. Van Gool, E. Pauwels and A. Oosterlinck, "Determination of optical flow and its discontinuities using non-linear diffusion". *In 3rd European Conference on Computer Vision*, vol. 2, pp. 295–304, 1994.

[21] P. Anandan, “A Computational Framework and an algorithm for the measurement of visual motion”. *International Journal of Computer Vision*, no. 2, pp. 283-310, 1989.

[22] B. Buxton and H. Buxton, "Computation of optical flow from the motion of edge features in image sequences". *Image and Vision Computing*, no. 2, pp. 59-74, 1984.

[23] D. J. Fleet and A. D. Jepson, "Computation of component image velocity from local phase information". *International Journal of Computer Vision*, no. 5, pp. 77-104, 1990.

[24] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision". *Proc. of DARPA IU Workshop*, pp. 121-130, 1981.

[25] H. H. Nagel, "On the estimation of optical flow relations between different approaches and some new results". *Artificial Intelligence*, no. 33, pp. 299-324, 1987.

[26] D. J. Heeger, "Optical flow using spatiotemporal filters". *International Journal of Computer Vision*, no. 1, pp. 279-302, 1988.

[27] A. Singh, "An estimation-theoretic framework for image-flow computation". *Proc. of ICCV*, pp. 168-177, Osaka, 1990.

[28] A. M. Waxman, J. Wuand and F. Bergholm, "Convected activation profiles and receptive fields for real time measurement of short range visual motion". *Proc. of IEEE CVPR*, pp. 717-723, Ann Arbor, 1988.

[29] S. Uras, F. Girosi, A. Verri and V. Torre, "A computational approach to motion perception". *Biological Cybernetics*, no. 60, pp. 79-97, 1988.

[30] H. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover". *Technical Report CMU-RI-TR-3 Carnegie-Mellon University, Robotics Institute*, 1980.

[31] C. Harris and M. Stephens, "A combined corner and edge detector". *Proc. of the 4th Alvey Vision Conference*. pp. 147–151, 1988.

[32] J. Shi and C. Tomasi, "Good features to track". *In IEEE Conference on Computer Vision and Pattern Recognition, Seattle*, June 1994.

[33] D. G. Lowe, "Object recognition from local scale-invariant features". *International Conference on Computer Vision*, Corfu, Greece, pp. 1150-1157, 1999.

[34] I. Laptev and T. Lindeberg, "Space-Time interest points". *Proc. of ICCV'03*, Nice, France, pp. 432-439, 2003.

[35] R. Chaudry, A. Ravichandran, G. Hager and R. Vidal, "Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions". *Proc. of CVPR*, Miami, US, 2009.

[36] A. Efros, A. Berg, G. Mori and J. Malik, "Recognizing action at a distance". *In IEEE International Conference on Computer Vision*, pp. 726–733, 2003.

[37] D. Tran and A. Sorokin, "Human activity recognition with metric learning". *In European Conference on Computer Vision*, 2008.

[38] Y. Yi and M. Lin, "Human action recognition with graph-based multiple-instance learning", *Pattern Recognition*, vol. 53, pp. 148-162, 2016.

[39] O. Fatemi, S. Zhang and S. Panchanathan, "Optical flow based model for scene cut detection". *Canadian Conference on Electrical and Computer Engineering*, 1996.

[40] W. Xiong, J. Chung and M. Lee, "Efficient scene change detection and camera motion annotation for video classification". *Computer Vision and Image Understanding*, vol. 71, issue 2, pp. 166-181, 1998.

[41] A. Kowdle and T. Chen, "Learning to Segment a Video to Clips Based on Scene and Camera Motion". *Proc. of ECCV*, pp. 272-286, Italy, 2012.

[42] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253, 2007.

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

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

[45] J. C. Niebles and L. Fei-Fei, "A Hierarchical Model of Shape and Appearance for Human Action Classification". *In IEEE International Conference on Computer Vision and Pattern Recognition*, 2007.

[46] V. Chasanis, A. Likas and N. Galatsanos, "Simultaneous detection of abrupt cuts and dissolves in videos using support vector machines". *Pattern Recognition Letters*, 2009.

[47] T. Barbu, "Novel automatic video cut detection technique using Gabor filtering". *Computers and Electrical Engineering*, 2009.

[48] A. Whitehead, J. Bose and R. Laganière, "Feature based cut detection with automatic threshold selection". *International Conference on Image and Video Retrieval*, Dublin, Ireland, pp. 410-418, July 2004.

[49] D. Roghayeh, Dadashi, R. Hamidreza and K. Rashidy, "AVCD-FRA: A novel solution to automatic video cut detection using fuzzy-rule-based approach". *Computer Vision and Image Understanding*, 2013.

[50] W. Effelsberg, "MOCA project". *Jahrestagunug University of Mannheim*; 1998.

[51] J. Lee, S-J Kim and C. S. Lee, "Effective Scene Change Detection by Using Statistical Analysis of Optical Flows". *Applied Mathematics and Information Sciences*, no. 6, pp. 177S-183S, 2012.

[52] P. Weinzaepfel, J. Revaud, Z. Harchaoui and C. Schmid, "Learning to Detect Motion Boundaries". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2578-2586, Boston, MA, USA, 2015.

[53] P. Sundberg, T. Brox, M. Maire, P. Arbelaez and J. Malik, "Occlusion boundary detection and figure/ground assignment from optical flow". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2233-2240, Colorado Springs, CO, USA, 2011.

[54] H. Fu, C. Wang, D. Tao and M. J. Black, "Occlusion Boundary Detection via Deep Exploration of Context". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 241-250, 2016.

[55] S. Phan, T. D. Ngo, V. Lam, S. Tran, D. D. Le, D. A. Duong and S. Satoh, "Multimedia Event Detection Using Segment-Based Approach for Motion Feature". *Journal of Signal Processing Systems*, vol. 74, no. 1, pp. 19-31, 2014.

[56] S. S. Kumar and M. John, "Human activity recognition using optical flow based feature set". *IEEE International Carnahan Conference on Security Technology (ICCST)*, pp. 1-5, Orlando, FL, 2016.

[57] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy and L. Fei-Fei, "Detecting Events and Key Actors in Multi-Person Videos". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3043-3053, 2016.

[58] G. Pass and R. Zabih, “Comparing images using joint histograms”. *Multimedia Systems*, 7(3): 234-240, 1999

[59] S. Little, I. Jargalsaikhan, K. Clawson, M. Nieto, H. Li, C. Direkoglu, N. E. O'Connor, A. F. Smeaton, B. Scotney, H. Wang and J. Liu, "An information retrieval approach to identifying infrequent events in surveillance video". *Proc. of the 3rd ACM Conference on International Conference on Multimedia Retrieval* (*ICMR*), pp. 223–230, ACM, New York, NY, USA, 2013.

[60] T. Wang and H. Snoussi, "Histograms of optical flow orientation for abnormal events detection". 2013 *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS)*, pp. 45–52, Clearwater, FL, USA, 2013.

[61] H. Wang and C. Schmid, "Action recognition with improved trajectories". *The IEEE International Conference on Computer Vision (ICCV)*, pp. 3551-3558, 2013.

[62] K. Guo, P. Ishwar and J. Konrad, "Action Recognition Using Sparse Representation on Covariance Manifolds of Optical Flow". *IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 188-195, Boston, MA, 2010.

[63] P. Matikainen, M. Hebert and R. Sukthankar, "Trajectons: Action recognition through the motion analysis of tracked features". *IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pp. 514-521, Kyoto, 2009.

[64] X. Li, "HMM based action recognition using oriented histograms of optical flow field". *Electronics Letters*, vol. 43, no. 10, pp. 560-561, 2007.

[65] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features". *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, Anchorage, AK, 2008.

[66] S. Ali and M. Shah, "Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 288-303, Feb. 2010.

[67] P. C. Ribeiro and J. S. Victor, "Human Activity Recognition from Video: modeling, feature selection and classification architecture". *International Workshop on Human Activity Recognition and Modelling*, Oxford, UK, 2005.

[68] A. Mekonnen, S. K. Selvi, V. S. Kumar and B. Tesfaye "Optical flow based Ethiopian traditional dance video classification system". *Computer Science and Telecommunications*, no.1(47), 2016.

[69] J. R. R. Uijlings, I. C. Duta, N. Rostamzadeh and N. Sebe, "Realtime Video Classification using Dense HOF/HOG". *In Proc. of International Conference on Multimedia Retrieval*, ACM, New York, NY, USA, 2014.

[70] I. C. Duta, "Histograms of Motion Gradients for real-time video classification". *14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1-6., Bucharest, 2016.

[71] M. H. Kolekar and D. P. Dash, "Hidden Markov Model based human activity recognition using shape and optical flow based features". *IEEE Region 10 Conference (TENCON)*, pp. 393-397, Singapore, 2016.

[72] H. Wang, D. Oneata, J. Verbeek and C. Schmid, "A Robust and Efficient Video Representation for Action Recognition". *International Journal of Computer Vision*, volume 119, Issue 3, pp. 219–238, 2016.

[73] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York: Dover Publications, 1970.

[74] G. B. Thomas and R. L. Finney, *Calculus and Analytic Geometry (9th ed.)*, Addison Wesley, 1996.

[75] M. Greenberg, *Advanced Engineering Mathematics (2nd ed.)*, Prentice Hall, 1998.

[76] B. K. P. Horn, *Robot Vision*. Cambridge, Mass.: MIT Press, 1986.

[77] R. Lienhart, “Reliable Transition Detection in Videos: A Survey and Practitioners Guide”. *International Journal of Image and Graphics*, 1(3): 469–486, 2001.

[78] A. Eweiwi, M. S. Cheema and C. Bauckhage, “Action recognition in still images by learning spatial interest regions from videos”. *Pattern Recognition Letters*, vol. 51, pp. 8-15, 2015.

[79] E. Park, X. Han, T. L. Berg and A. C. Berg, "Combining multiple sources of knowledge in deep CNNs for action recognition". *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1-8, Lake Placid, NY, 2016.

[80] L. Liu, L. Shao, X. Li and K. Lu, "Learning Spatio-Temporal Representations for Action Recognition: A Genetic Programming Approach". *In IEEE Transactions on Cybernetics*, vol. 46, no. 1, pp. 158-170, 2016

APPENDIX A

ENVIRONMENT AND LIBRARIES

The implementation about video processing is generally done by using C++ programming language in Linux environment.

For the state-of-the-art algorithms in Computer Vision, OpenCV library is used. Open CV was very useful especially on optical flow calculations with its broad library support.

Open CV also provides library support for feature extraction and tracking. So, it helps us extracting the features used in optical flow calculations.

In action and cut classification, the implementation is done by using Java programming language in Linux environment. Linux bash scripting is also used in some cases for automating processes.

For SVM classification, LIBSVM library is selected. Its java based package is utilized with various configurations.

APPENDIX B

CLASSIFIED ACTION SAMPLES

The figures below - starting with Figure B.1 and ending with B.10 - include frames from the Hollywood Human Actions dataset and calculated optical flow representations for those frames' transitions.

Each figure representing a frame sequence has its optical flow representation occurred in the transition of its frames. Each two consecutive figures can be considered as sample pairs for each action type.

The frame samples are selected for giving an insight about the action samples and optical flow characteristics. Five types of actions are shown in the figures:

- StandUp
- SitDown
- HandShake
- Hug
- SitUp

Figure B.1: Sample consecutive frames inside an action classified as StandUp

Figure B.2: Optical flow vectors for the sample consecutive frames of the StandUp action

Figure B.3: Sample consecutive frames inside an action classified as SitDown

Figure B.4: Optical flow vectors for the sample consecutive frames of the SitDown action

Figure B.5: Sample consecutive frames inside an action classified as HandShake

Figure B.6: Optical flow vectors for the sample consecutive frames of the HandShake action

Figure B.7: Sample consecutive frames inside an action classified as Hug

Figure B.8: Optical flow vectors for the sample consecutive frames of the Hug action

Figure B.9: Sample consecutive - three - frames inside an action classified as SitUp

Figure B.10: Optical flow vectors for the sample consecutive - three - frames of the SitUp action

APPENDIX C

DATA SETS

Table C.1: Hollywood Human Actions dataset

Action	**Number of Actions**		
	Training Set	**Test Set**	**Total**
AnswerPhone	22	23	45
GetOutCar	13	13	26
HandShake	20	19	39
HugPerson	22	22	44
Kiss	45	49	94
SitDown	44	27	71
SitUp	11	10	21
StandUp	42	48	90

Table C.2: Weizmann dataset

Action	**Number of Videos**
walk	10
run	10
jump	9
side	9
bend	9
wave (O.H.)	9
wave (T.H.)	9
pjump	9
jack	9
skip	10

Table C.3: Video Segmentation Project in Carleton University dataset

Set	**Number of Instances**		
	Cut	**Noncut**	**Total**
A	7	643	650
B	8	951	959
C	54	1565	1619
D	34	2598	2632
E	30	506	536
F	0	236	236
G	18	482	500
H	38	5095	5133
I	4	475	479
J	87	786	873

VITA

PERSONAL INFORMATION

Surname, Name: Akpınar, Samet
Nationality: Turkish (TC)
Date and Place of Birth: October 14, 1982, Konya
Marital Status: Married

EDUCATION

Degree	Institution	Year of Grad.
M.S.	Computer Engineering Dept., METU	2007
B.S.	Computer Engineering Dept., METU	2005
High School	Konya Meram Anadolu Lisesi	2000

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2010 - Present	Central Bank of Rep. of Turkey	Senior Software Engineer
2007 - 2010	Computer Eng. Dept., METU	Teaching Assistant
2006 - 2007	Computer Eng. Dept., METU	Research Assistant
2005 - 2006	Cybersoft	Software Engineer

PUBLICATIONS

Journal Publications

S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Çiçekli and F. N. Alpaslan, "An Ontology-Based Retrieval System Using Semantic Indexing". *Information Systems*, Volume 37, Issue 4, pp. 294-305, 2012.

Book Chapters

S. Akpınar and F. N. Alpaslan, "Optical flow-based representation for video action detection". *Emerging Trends in Image Processing, Computer Vision and Pattern Recognition*, Chapter 21, pp. 331-351, 2015.

Conference Publications

S. Kara, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Çiçekli and F. N. Alpaslan, "An ontology-based retrieval system using semantic indexing". *2010 IEEE 26th International Conference on Data Engineering Workshops (ICDEW 2010)*, pp. 197-202, Long Beach, CA, 2010.

M. Boyar, Ö. Alan, S. Akpınar, O. Sabuncu, N. K. Çiçekli and F. N. Alpaslan, "Event boundary detection using audio-visual features and web-casting texts with imprecise time information," *2010 IEEE International Conference on Multimedia and Expo*, pp. 578-583, Suntec City, 2010.

D. Tunaoğlu, Ö. Alan, O. Sabuncu, S. Akpınar, N. K. Çiçekli and F. N. Alpaslan, "Event Extraction from Turkish Football Web-casting Texts Using Hand-crafted Templates," *2009 IEEE International Conference on Semantic Computing*, pp. 466-472, Berkeley, CA, 2009.

O. Alan, S. Akpınar, O. Sabuncu, N. Çiçekli and F. Alpaslan, "Ontological video annotation and querying system for soccer games," *2008 23rd International Symposium on Computer and Information Sciences*, pp. 1-6, Istanbul, 2008.