MOTION ANNOTATION IN COMPLEX VIDEO DATASETS

Muhammad Habib Mahmood

PhD Thesis

Motion Annotation in Complex Video Datasets

Muhammad Habib Mahmood

2018

PhD Thesis

Motion Annotation in Complex Video Datasets

Muhammad Habib Mahmood

2018

DOCTORAL PROGRAM IN TECHNOLOGY

Supervised by: Dr. Arnau Oliver

Work submitted to the University of Girona in partial fulfillment of the requirements for the degree of Doctor of Philosophy

In the loving memory of my father...

He was, still is, and will always be, my greatest inspiration.

Acknowledgements

First and foremost, I thank the Almighty for the motivation, belief, circumstances and support that helped me reach this culmination of my research. I believe that courage and inspiration are the essential ingredients that make up and sustain any endeavor. I am lucky that during my doctoral studies I have always had a source of both, and for that, I am sincerely grateful to the pillars of my research work, Dr. Xavier Lladó and Prof. Joaquim Salvi. Without their brilliant ideas and continuous support, this work would simply not be possible. I am also indebted to my supervisor Dr. Arnau Oliver, who rescued my thesis when I needed someone the most.

This endeavor would not have been possible without my family that has always been there for me despite the overwhelming distances between us. Their love and encouragement helped me immensely in the completion of this thesis. My deepest gratitude to my mother, her unswerving support remains a cornerstone in my life. During the course of my doctoral studies, I lost my father, but still feel his silent encouragement, telling me to keep on going in the pursuit of my life's goals. My deep felt thanks to my brothers, sister, sister-in-law, grandparents, nephews, nieces and, especially my parents-in-law. My father-in-law especially remained a hidden support for us throughout.

My special thanks go to my friends many of whom are fellow PhD students and know first-hand what it takes to be one. They were there to help me out whenever I needed it the most: Konstantin, Guillaume, Mojdeh, Shihav, Sonia, Matheiu, Hamed, Juan, Richa, Mostafa, Sergi, Masi, Ricard P., Nuno, Josep, Eloy and Usman.

Also, I would like to thank my colleagues: Ferran and Yago, for their time and invaluable suggestions; Josep, Ricard P. and Ricard C., for taking the trouble to help me with all my technical problems; and Joseta, Mireia, and Montse, for keeping my back in the ruthless world of paperwork. My gratitude, as well, to the anonymous reviewers and the members of the defense panel for evaluating my work. Besides, this thesis largely owes its completion to the AGAUR FI-DGR 2012 grant provided by the Autonomous Government of Catalonia. I am grateful and honored to be among its holders.

Last but never least, I wish to thank my wife, Sarah, and my sons, Hamza and Qasim, for standing beside me throughout my PhD. Sarah's unwavering faith in me, even at the times when I felt that I cannot move forward, her patience and love encouraged me to not give up and accomplish what I had undertaken.

Publications

This thesis resulted in the publication of the following research articles:

- M. H. Mahmood, Y. Diéz, A. Oliver, J. Salvi and X. Lladó, "Motion Region Annotation for Complex Videos via Label Propagation Across Occluders," *Information Sciences*, 2018, under review. [JCR CSA IF:4.832 Q1(38/140)]

- M. H. Mahmood, Y. Diéz, J. Salvi and X. Lladó, "A Collection of Challenging Motion Segmentation Benchmark Datasets," *Pattern Recognition*, 2017, 61, 1–14. [JCR CSAI IF:4.582 Q1(15/130)]

- M. H. Mahmood, J. Salvi and X. Lladó, "Semiautomatic tool for Motion Annotation iin Complex Videos," *Electronics Letters*, 2016, 52-8, 602–604. [JCR EEE IF:0.930 Q3(147/249)]

- M. H. Mahmood, L. Zapella, Y. Diéz, J. Salvi and X. Lladó, "A New Trajectory based Motion Segmentation Dataset," *Proceedings of the 7th Iberian Conference on Pattern Recognition and Image Analysis*, 2015, 463–470.

Abbreviations

A-ASA	Automatic-Adaptive subspace Affinity
ALC	Agglomerative Lossy Compression
CPD	Coherent Point Drift
CNN	Convolutional Neural Network
ELSA	Enhanced Local Subspace Affinity
EM	Expectation Maximization
FN	False Negatives
FoV	Field of View
FP	False Positives
GPCA	Generalized Principal Component Analysis
GPU	Graphic Processing Unit
GT	Ground Truth labels
GUI	Graphical User Interface
HD	High Definition
KLT	Karhunen-Loève Transform
k**NN**	k-Nearest Neighbors
LDOF	Large Displacement Optical Flow

ABBREVIATIONS

LRR	Low Rank Representation
LS3C	Latent Space Sparse Subspace Clustering
ML	Misclassified motion Labels
MS	Motion Segmentation
OB	Och Brox
OM	Overall Misclassification
RGB	Red Green Blue (color space)
ROI	Region of Interest
SIFT	Scale-Invariant Feature Transform
SL	Segmentation Labels
SM	Motion Separation
SSC	Sparse Subspace Clustering
SURF	Speeded Up Robust Features
TAT	Trajectory Annotation Tool
TP	True Positive

List of Figures

List of Tables

List of Algorithms

Contents

Resum

L'anàlisi crític i numèric de les diferents metodologies de la visió per ordinador acaba depenent en gran mesura dels datasets sobre els quals aquestes metodologies es posen a prova. Qualsevol conjunt de dades (d'imatges) és tant bo com la diversitat de les dades que existeixen en el propi problema. La segmentació dels moviments, motion segmentation en anglès, és un pas de processament previ de la visió per ordinador crític en diferents aplicacions i, en canvi, els conjunts de dades públics presenten certes limitacions. Algunes bases de dades no estan al dia amb les exigències modernes de la longitud dels vídeos i el nombre de moviments, mentre que d'altres no tenen una bona avaluació manual, fet necessari per a la posterior comparació d'algorismes. En aquest treball presentem una col·lecció de datasets multi-facètics de segmentació del moviments que contenen anotacions manuals tant de les trajectòries dels moviments com de les regions afectades. Aquests conjunts de dades presenten seqüències quotidianes de diferent temporització, però amb un alt nombre de moviments i de fotogrames per seqüència, així com distorsions i oclusions. L'anotació manual es proporciona en totes les imatges de totes les seqüències. Es proporciona també una avaluació exhaustiva, i de referència, dels algoritmes de segmentació del moviment amb tecnologia d'última generació, útil per a establir la dificultat del problema i contribuir també a un nou punt de partida.

L'anotació manual dels datasets de segmentació del moviment en escenes quotidianes del dia a dia és una tasca difícil i desafiant. La comunitat científica encara manca d'una eina d'anotació estàndard per a aquest tipus de dades, fet que fa que sigui un camp d'investigació encara molt obert. En aquesta tesi doctoral proposem una eina d'anotació de trajectòries complexes, proporcionant una plataforma pública i única per crear i reforçar les bases de dades de segmentació dels moviments. La intuïtiva interfície d'usuari permet refinar un resultat inicial de la segmentació del moviments obtingut automàticament per tal d'obtenir una avaluació manual dels moviments al llarg de tots els fotogrames d'una seqüència donada. En vídeos llargs amb múltiples moviments rígids i no rígids amb oclusions completes i distorsions reals, la nostra eina facilita la ràpida anotació dels moviments.

La característica del moviment és fonamental en l'anàlisi dels moviments dels objectes, procés clau per a la detecció dels moviments i posterior segmentació d'aquests. Aquesta tasca de pre-processament és un pilar clau per a aplicacions tan diverses com ara el reconeixement, la similitud, l'estimació, etc, tant de moviments com d'objectes. Per tal d'idear un algoritme robust per a l'anàlisi de moviment, és imprescindible comptar amb un ampli conjunt de dades per avaluar-ne el rendiment. La principal limitació en l'obtenció d'aquest tipus de conjunts de dades és la creació de la anotacions del moviment en les imatges, ja que cada objecte en moviment pot abastar diversos fotogrames, i pot veure's sotmès a canvis en la grandària, la il·luminació i l'angle de vista. A més dels canvis òptics, l'objecte pot patir oclusions per objectes, tant estàtics com en moviment. El repte augmenta encara més de dificultat quan el vídeo que s'està processant ha estat obtingut per una càmera en moviment. En aquesta tesi, també abordem aquests tipus de vídeos, proporcionant mètodes per a facilitar-ne l'anotació manual. A partir d'una anotació manual mínima que consisteix en crear una màscara sobre l'objecte en qüestió, som capaços de propagar la màscara (etiqueta) a tots els demés fotogrames. Quan un objecte pateix oclusions per part d'un objecte estàtic o en moviment, les màscares es mantenen pels dos objectes i es defineix una ordenació de les màscares en funció de la profunditat d'aquests. A més a més, també s'han dissenyat un conjunt de mesures innovadores per tal d'avaluar el rendiment dels algoritmes de segmentació dels moviments. Els resultats mostren que el nostre enfocament en cascada proporciona resultats satisfactoris en una gran varietat de seqüències de vídeo contenint escenes quotidianes.

Resumen

El análisis crítico y numérico de las diferentes metodologías de la visión por ordenador termina dependiendo en gran medida de los datasets sobre los que estas metodologías se ponen a prueba. Cualquier conjunto de datos (de imágenes) es tan bueno como la diversidad de los datos que existen en el propio problema. La segmentación de los movimientos, motion segmentation en inglés, es un paso de procesamiento previo de la visión por ordenador crítico en diferentes aplicaciones y, en cambio, los conjuntos de datos públicos presentan ciertas limitaciones. Algunas bases de datos no están al día con las exigencias modernas de la longitud de trama y el número de movimientos, mientras que otras no tienen una buena evaluación manual, clave para la posterior comparación de algoritmos. En este trabajo presentamos una colección de datasets multifacéticos de segmentación del movimiento que contienen anotaciones manuales tanto de las trayectorias de los movimientos como de las regiones afectadas. Estos conjuntos de datos presentan secuencias cotidianas de diferente temporización, pero con un alto número de movimientos y de imágenes por secuencia, así como distorsiones y oclusiones. La anotación manual se proporciona en todos los fotogramas de todas las secuencias. Se proporciona también una evaluación exhaustiva, y de referencia, de los algoritmos de segmentación del movimiento con tecnología de última generación, útil para establecer la dificultad del problema y contribuir también a un nuevo punto de partida.

La anotación manual de los datasets de segmentación del movimiento en escenas cotidianas del día a día es una tarea difícil y desafiante. La comunidad científica todavía carece de una herramienta de anotación estándar para este tipo de datos, lo que hace que sea un campo de investigación todavía muy abierto. En esta tesis doctoral proponemos una herramienta de anotación de trayectorias complejas, proporcionando una plataforma pública y única para crear y reforzar las bases de datos de segmentación de los movimientos. La intuitiva interfaz de usuario permite refinar un resultado inicial de la segmentación del movimiento obtenido automáticamente para obtener una evaluación manual de los movimientos a lo largo de todos los fotogramas de una secuencia dada. En videos largos con múltiples movimientos rígidos y no rígidos con oclusiones completas y distorsiones reales, nuestra herramienta facilita la rápida anotación de los movimientos.

La característica del movimiento es fundamental en el análisis de los movimientos de los objetos, proceso clave para la detección de los movimientos y posterior segmentación de los mismos. Esta tarea de pre-procesamiento es un pilar clave para aplicaciones tan diversas como el reconocimiento, la similitud, la estimación, etc, tanto de movimientos como de objetos. Para idear un algoritmo robusto para el análisis de movimiento, es imprescindible contar con un amplio conjunto de datos para evaluar su rendimiento. La principal limitación en la obtención de este tipo de conjuntos de datos es la creación de la anotación del movimiento en las imágenes, ya que cada objeto en movimiento puede abarcar varios fotogramas, y puede verse sometido a cambios en el tamaño, la iluminación y el ángulo de vista. Además de los cambios ópticos, el objeto puede sufrir oclusiones por objetos, tanto estáticos como en movimiento. El reto aumenta aún más de dificultad cuando el vídeo que se está procesando ha sido obtenido por una cámara en movimiento. En esta tesis, también abordamos este tipo de vídeos, proporcionando métodos para facilitar en éstos la anotación manual. A partir de una anotación manual mínima que consiste en crear una máscara sobre el objeto en cuestión, somos capaces de propagar la máscara (etiqueta) a todos los demás fotogramas. Cuando un objeto sufre oclusiones por parte de un objeto estático o en movimiento, las máscaras se mantienen para los dos objetos y se define una ordenación de las máscaras en función de la profundidad de los mismos. Además, también se han diseñado un conjunto innovador de medidas para evaluar el rendimiento de los algoritmos de segmentación de los movimientos. Los resultados muestran que nuestro enfoque en cascada proporciona resultados exitosos en una gran variedad de secuencias de vídeo conteniendo escenas cotidianas.

Abstract

An in-depth analysis of computer vision methodologies is greatly dependent on the benchmarks they are tested upon. Any dataset is as good as the diversity of the true nature of the problem enclosed in it. Motion segmentation is a preprocessing step in computer vision whose publicly available datasets have certain limitations. Some databases are not up-to-date with modern requirements of frame length and number of motions, and others do not have ample ground-truth in them. In this paper, we present a collection of diverse multifaceted motion segmentation benchmarks containing trajectory- and region-based ground-truth. These datasets enclose real-life long and short sequences, with increased number of motions and frames per sequence, and also real distortions with missing data. The ground-truth is provided on all the frames of all the sequences. A comprehensive benchmark evaluation of the state-of-the-art motion segmentation algorithms is provided to establish the difficulty of the problem and to also contribute a starting point.

Ground-truth annotation on motion segmentation datasets of arbitrary real-life videos is a difficult and challenging task. The research community lacks a standard annotation tool for such datasets, which makes it an open research field. We propose in this PhD thesis, an annotation tool for trajectories in complex videos, which provides a publicly available platform to create and reinforce motion segmentation datasets. The user friendly interface allows to refine an initial automatic segmentation result to produce ground-truth annotation on all the motions of all the frames of a given sequence. In long videos with multiple rigid/nonrigid motions containing complete occlusion and real distortions, our tool facilitates rapid annotation of motion in a semi-automatic way.

Motion cue is pivotal in moving object analysis, which is the root for motion segmentation and detection. These preprocessing tasks are building blocks for several applications such as recognition, matching, estimation, etc. To devise a robust algorithm for motion analysis, it is imperative to have a comprehensive dataset to evaluate an algorithm's performance. The main limitation in making these kind of datasets is the creation of ground-truth annotation of motion, as each moving object might span over multiple frames with changes in size, illumination and angle of view. Besides the optical changes, the object can undergo occlusion by static or moving occluders. The challenge increases many-fold when the video being processed is captured by a moving camera. In this thesis, we also tackle the task of providing ground-truth annotation on motion regions in videos captured from a moving camera. With minimal man-

ual annotation of an object mask, we are able to propagate the label mask in all the frames. Object label correction based on static and moving occluder is also performed by applying occluder mask tracking for a given depth ordering. A motion annotation dataset is also proposed to evaluate the algorithm performance. The results show that our cascaded-naive approach provides successful results in a variety of video sequences.

'Never confuse motion with action'.

Benjamin Franklin

Chapter 1

Introduction

The advances being done in computer vision are greatly dependent on the characteristics of the data bank of images and videos which algorithms use as benchmark. If the data encapsulates the true nature of the desired problem, then the proposed solution can be rigorously tested, results can be repeatable, standardization of evaluation measures can be achieved, and new algorithms can be adequately compared. However, if the bank of data does not envelop the correct nature of the task in question, or if a trivial subset of the problem is captured which is not representative of the actual problem, then the solutions proposed might be limited in nature, unstable in results, and thus might hamper further research. Due to these reasons, nowadays, the importance of datasets is immense as they are shaping the way forward for computer vision algorithms.

1.1 Motion cue

Motion cue is pivotal in moving object analysis, which is the root for motion segmentation and detection. These preprocessing tasks are building blocks for several applications such as recognition, matching, estimation, etc. Motion analysis is a prerequisite in video analysis with its applications in computer vision ranging from surveillance [1–4], crowd estimation [5–10] to gesture recognition [11,12], video object segmentation [13–18], behavior analysis [19, 20] and anomaly detection [21–23]. An objective analysis of moving objects can be carried out, when motion is accurately detected and segmented as a prior. In the state-of-the-art of computer vision, precise and robust algorithms, which can work in the presence of occluders and distortions, while the acquisition is done from a moving camera, shown in Fig. 1.1 are still elusive. Therefore,

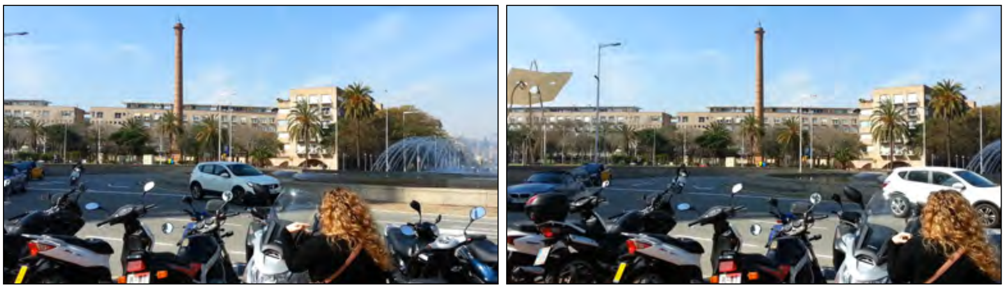


Figure 1.1: An example of a real video of a natural outdoor scene. **Left:** A bike and three cars are in motion in the shadowy region in the field of view. A person is walking past in front of the moving camera. **Right:** In subsequent frames, the white car comes out of the shadow goes under occlusion behind the moving person and the static occluder objects.

research to find a solution of these tasks is still an open field.

1.2 Research motivation

In many computer vision tasks the decomposition of the video into moving objects and background is the first fundamental step. It is an essential building block for robotics, inspection, metrology, video surveillance, video indexing, traffic monitoring and many other applications.

1.2.1 Motion segmentation problem

Motion Segmentation (MS) is one such research area, in which temporally continuous set of frames, called a sequence, are processed to provide a unique label to every motion present in those frames. It is a preprocessing step for several computer vision problems, i.e. semantic segmentation, crowd estimation, surveillance [3, 4, 9, 24]. MS can be performed on precomputed set of sparse trajectories of a sequence, or it can be directly applied on a desired set of frames to get motion region labels.

Researchers have shown particular interest towards MS in recent years [25–29]. The problem has been addressed in a variety of approaches. The question was initially introduced as a subspace affinity problem in [30]. This approach was then extended as a sparse subspace clustering (SSC) problem [31] by proposing the use of sparse representation of data drawn from multiple low-dimensional linear or affine subspaces. Similarly, a low-rank representation (LRR)[32, 33] to segment data drawn from a union of multiple linear or affine subspaces was proposed. Unlike the sparse representation SSC, which computes the individual sparsest representation of each data vector, LRR aimed to find the lowest-rank representation of a collection of vectors jointly. Another method with a new interpretation to extract the rank of trajectory matrix and an improved affinity measure was also recently proposed [34–36]. The approach in [37] introduced a temporal subspace clustering method for unsupervised segmentation of human motion by designing a temporal Laplacian regularization function to model the sequential information in time. This method had limited applicability in general MS problems.

The performance of all these algorithms can only be compared with standardized benchmark datasets, which are accepted by the community. The making of such datasets gives rise to another subproblem in MS.

1.2.2 Datasets for motion segmentation evaluation

In the huge amount of video data that is captured by hand-held devices these days, moving objects exhibit a variety of characteristics. MS datasets must be representative of these characteristic attributes of motion to capture the diversity of the problem.

The main limitations of the state-of-the-art of MS datasets, as seen in Fig. 1.2, are that they are made up of sets of videos, where the use of synthetic sequences, like passively moving checker boards or other static objects, still prevail [38]. The motions in these datasets are captured in short sequences, with little or no background change. The number of motions per sequence are few, and even a video shot capturing 4 or 5 moving objects does not exhibit considerable amount of movement as is present in daily-life natural scenes. The results of algorithms are put forth on these databases because of the absence of a challenging, diverse dataset. In some cases, improved results are presented on less representative databases which does not help the cause. There are other small datasets, like Extended-Yale benchmark [40], SegTrack [41] and Cambridge CamShift dataset [42], which motion segmentation community has used to present their results. They are limited and not widely used because of fewer sequences or a smaller number of motions. A recently proposed dataset [39] tried to overcome these limitations, but the ground-truth is provided only on 5% of the frames. Therefore, the need for a more diverse and challenging motion segmentation database, with which the borders of research in MS algorithms can be expanded, still persists.

1.2.3 Motion annotation in video sequences

The limitations prevailing in annotated moving objects' datasets are restricting the development of effective motion analysis tools. The diversity and complexity of a real life

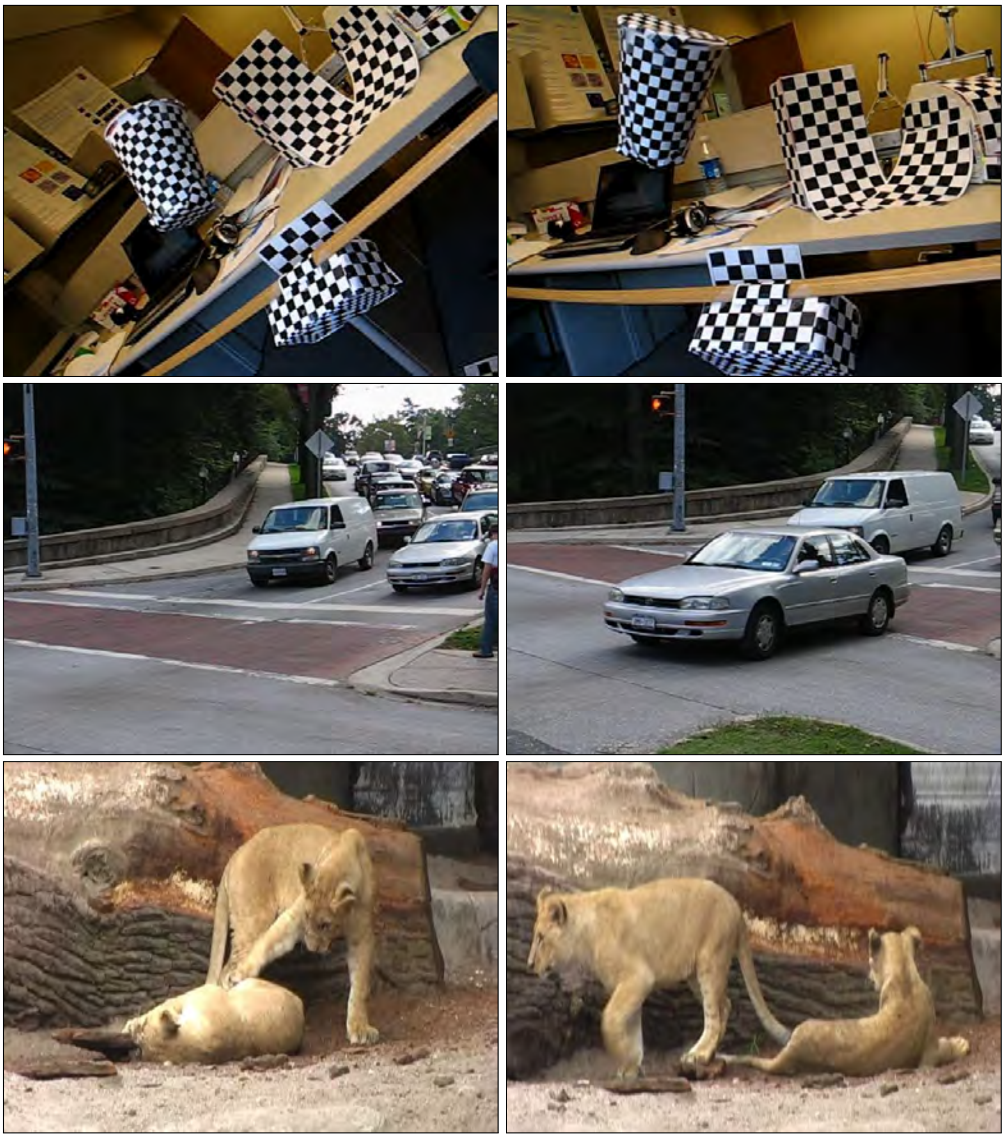


Figure 1.2: First and last frame of video shots from the state-of-the-art in MS datasets. **Top:** '1R2TCRT', total 42 frames in Hopkins [38]. **Center:** 'cars9', total 60 frames in Hopkins and FBMS [39] both. **Bottom:** 'lion02' total 416 frames in FBMS.

motion captured in a collection of video sequences, determines how representative the dataset is of the actual problem. If the annotated datasets encapsulate limited motion diversity, then the algorithms tested on them will also have limited applicability. On the other hand, if more complex motions are captured in a sequence for dataset formation, the dataset will become more representative but the task of correctly generating ground-truth motion label for each moving object in all the frames of a video sequence becomes increasingly cumbersome. Here, the problematic element is the expert-user annotation time, which increases many-fold as the captured motion becomes excessively complex.

An illustrative example is presented in Fig. 1.3, which shows the first, middle and last frames of two moving objects in a video shot, while they enter and leave the field of view. The white car in the left column remains unoccluded, relative change in size across all frames is minimal, the illumination remains generally homogeneous and no perspective distortion effect can be seen. On the other hand, the blue truck, present in the right column, enters the field of view with a small size due to being considerably deep in the scene with reference to the camera, experiences complete occlusion during the course of its motion, and exits the frame with an enlarged size, change in heading direction, variation in illumination and with perspective distortion. The expert-user annotation time for generating ground-truth on these two motion samples is radically different. While the annotation labels on the white car be provided with state-of-the-art label propagation algorithms, there is no modern, time efficient methodology or platform, to annotate the blue truck or such motions.

This limitation in label propagation can be looked into as a set of multiple sub-problems based on the complexity and variation in the object motion. The variants include a considerable change in size or illumination, partial or complete occlusion, static or moving occluder, multiple-appearance-and-disappearance in the field of view (FoV), perspective distortion, etc. Each variant, if tackled separately, with a unique approach, can yield improved results.

1.3 Objectives

The main aim of the thesis was to overcome the problems present in the state-of-the-art of MS datasets by proposing a modern MS benchmark dataset. The construction of datasets involves ground-truth annotation. Keeping this in view, we presented motion annotation tools, and also presented evaluation methods and sequences to test the tools' performance. The details of our objectives are;

1. In our work, we present our multifaceted diverse publicly available motion segmentation benchmark dataset of 39 long, and 312 short sequences with ground-truth available in all the frames of all the sequences. The ground-truth on 19 long, and 162 short sequences is provided as a trajectory label on all the tra-

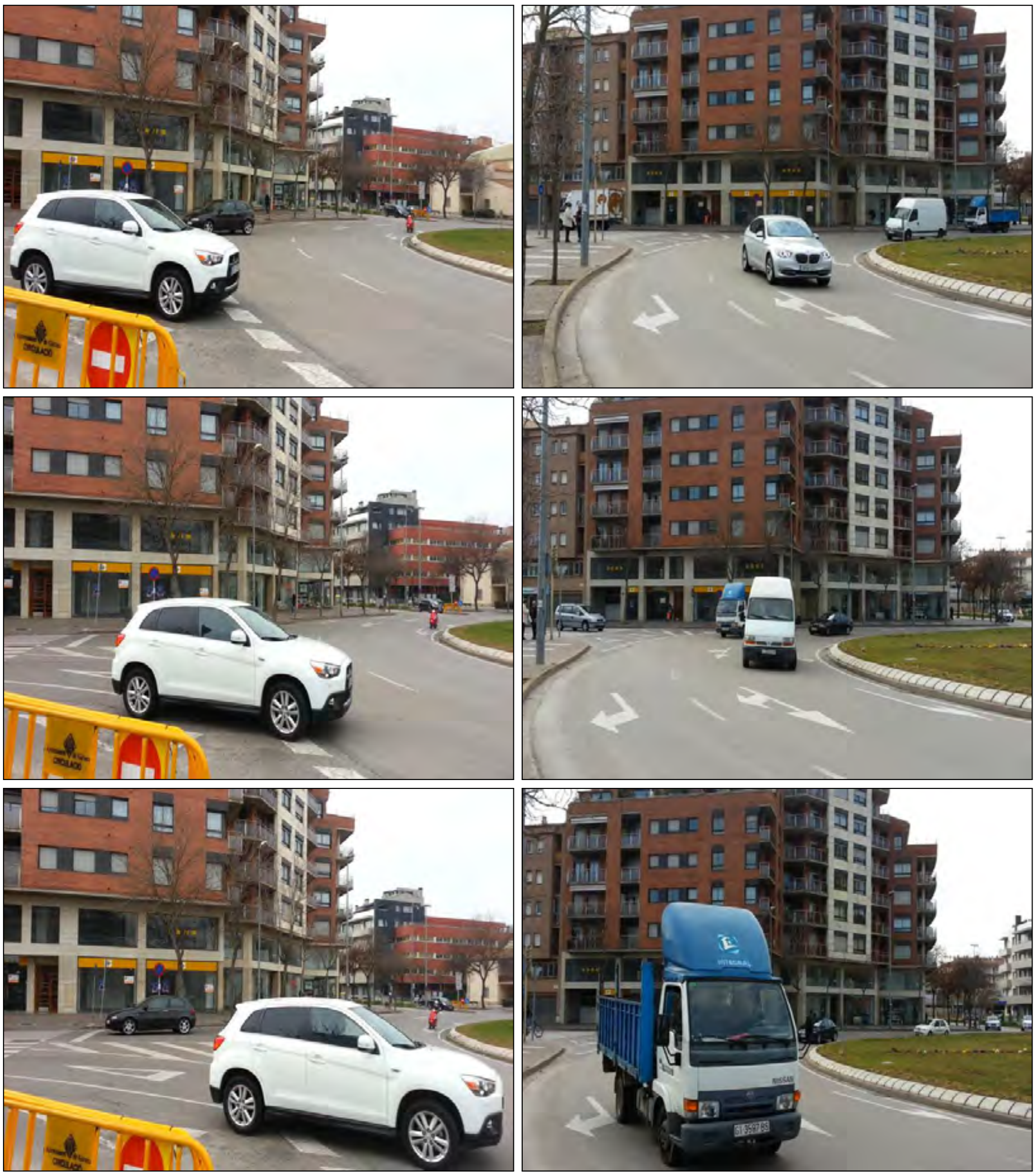


Figure 1.3: First, middle and last frame of a moving object, while entering and leaving the field of view in a video shot, in the Top, Middle and Bottom row of the figure. A comparison of the visible difference in moving objects' behavior as captured in real videos. **Top:** The white car in the left figure and the blue truck in the right figure enter the frame. Blue truck is small because of its depth with respect to the camera. **Middle:** The white car moves along without any distortion undergoing a mile depth change, while the blue truck experiences complete occlusion behind the white van. **Botom:** The white car continues the same behavior, while the blue is about to exit the frame with change in heading direction and perspective distortion. Annotation of objects undergoing such changes in object size, shape, heading direction and illumination is extremely difficult.

jectories of all the motions. And the ground-truth on 20 long, and 150 short sequences is provided as a region label on all the motion regions of all the sequences. The average number of motions per sequence is almost 11, and the average frame length is around 815 frames. With a goal to overcome all the limitations present in the state of the art, all the sequences contain a fusion of real noise and distortions. The captured characteristics of noise in the sequences are missing data, partial/complete/multiple occlusion, stopping motion, multiple appearance-disappearance of objects, perspective distortion, etc.

2. A further subset of 40 trajectory-based and 34 region-based short sequences with complete data are also provided to test algorithms that are unable to deal with missing data. Ochs-Brox (OB)[39] algorithm is used to provide a benchmark on long sequences as besides OB no other algorithm can be applied on these long sequences because of their complexity.

3. Moreover, on short sequences a benchmark analysis with six well-known state-of-the-art MS algorithms i.e. SSC[31], ALC[43], LRR[33], LS3C[44], ELSA[34], OB[39], is also presented for detailed quantitative and qualitative evaluation. The evaluation metrics encapsulate all the criteria used in the state-of-the-art datasets for an in-depth analysis. Moreover, a course of action on how to improve results on this database to produce robust algorithms is also suggested.

4. We propose here an annotation tool for trajectories in complex videos, which provides a publicly available platform to create and reinforce motion segmentation datasets. The user friendly interface allows to refine an initial automatic segmentation result to produce ground-truth annotation on all the motions of all the frames of a given sequence. In long videos with multiple rigid/nonrigid motions containing complete occlusion and real distortions, our tool facilitates rapid annotation of motion in a semi-automatic way.

5. In our work, we also propose a methodology, which efficiently utilizes the expert-user time to propagate labels on all moving objects in all the frames of a video sequence captured from a moving camera. With an existing platform [45], which propagates labels in situations with no occlusion nor distortions, our methodology is integrated to propagate labels across occlusions and its related distortions. The propagation result keeps the object shape intact with scale adjustment. We do so by using just two user labeled motion masks, the first and last frame of a sub-problem set. Utilizing the two masks, we perform object mask propagation

across all frames using maximal flow vector count, acquired through Large Displacement Optical Flow (LDOF) [46]. Concurrently, we take a static occluder shape input on a single frame from the user, to perform occluder mask tracking using keypoint descriptors (SURF features [47]) across all frames. With non-rigid point set registration [48,49] of the first frame mask onto the last frame, we perform object mask scale adjustment to improve the propagated object mask estimate.

6. To validate the performance of our approach, we carryout a quantitative and qualitative analysis of our algorithm on moving objects undergoing partial occlusion, where occluder is both static and moving, with sequences captured from a moving camera. In this regard, we put forth a 25 sequence occlusion/occluder dataset with moving objects going across static or moving occluder(s). On 20 static and 5 moving occluders, our results demonstrate that by splitting the motion annotation problem into sub-problem sets, the expert-user time is utilized with improved efficiency, maintaining accurate boundaries on the object annotations.

1.4 Thesis outline

This thesis describes the research work that resulted in the creation of a collection of benchmark motion segmentation datasets. The benchmarking task was performed with the state-of-the-art MS algorithms. The thesis also dwells on the motion annotation problem. A trajectory annotation tool for sparse motion trajectory labeling and a motion-region annotation framework are also presented. Prior to designing the proposals, we exhaustively studied the literature that exists in the field of MS.

Chapter 2 reports the findings of the literature review. A detailed insight of the state-of-the-art motion segmentation datasets and algorithms was presented. Moreover, tools and methods of motion annotation and label propagation as available in the literature were also studied.

Chapter 3 contains a description of the making of the trajectory- and region-based motion segmentation, long and short, datasets. The development of the benchmark based on recent algorithms, their working and the interpretation of results is also explained.

Chapter 4 presents the tools developed for motion annotation. The trajectory annotation tool needed to create the trajectory-based datasets is explained in detail with its supporting modalities. After that, the motion-region annotation framework is described along with its associated algorithms. A detailed quantitative and qualitative analysis is also given.

Chapter 5 concludes this thesis highlighting the contributions. This chapter also describes the limitations and suggests short term and long term future work.

Our review of the literature says this appears to be bigger than in the past.

Bob Dietz

Chapter 2

Literature review

Motion analysis has been of interest for researchers since long. The details encapsulated in a video are better judged, if the motion present in it is segmented and understood. This idea propelled the community to work on several different off shoots of motion analysis, which includes motion segmentation among others. Motion segmentation problem has it roots in the dataset it is being analyzed on. If the dataset is realistic, with real object motions and distortions, the segmentation proposal becomes that much more comprehensive. The basics of a MS dataset creation are embedded in motion annotation and label propagation. An automatic or semi-automatic motion annotation platform can facilitate the labeling of complex motions in real sequences of a dataset. This in turn would result in algorithms having a profound solution of the MS problem.

In this chapter, we present a detailed review of these three aspects of motion segmentation problem. In Section 2.1, a detailed overview of the state-of-the-art MS datasets has been presented. Following this overview, in Section 2.2, the recent MS segmentation methodologies, which have exhibited good results are presented. Section 2.3 contains a detailed note on the approaches taken by the community to solve the motion annotation problem. This section is followed by the concluding remarks in Section 2.4, including comments on the state-of-the-art of each problem.

2.1 Motion segmentation datasets

In this section, the state-of-the-art in MS datasets is presented. The attributes of each dataset are explained with an analysis along with the strengths and limitations.

The first segregation in MS datasets is the representation of motion annotation in

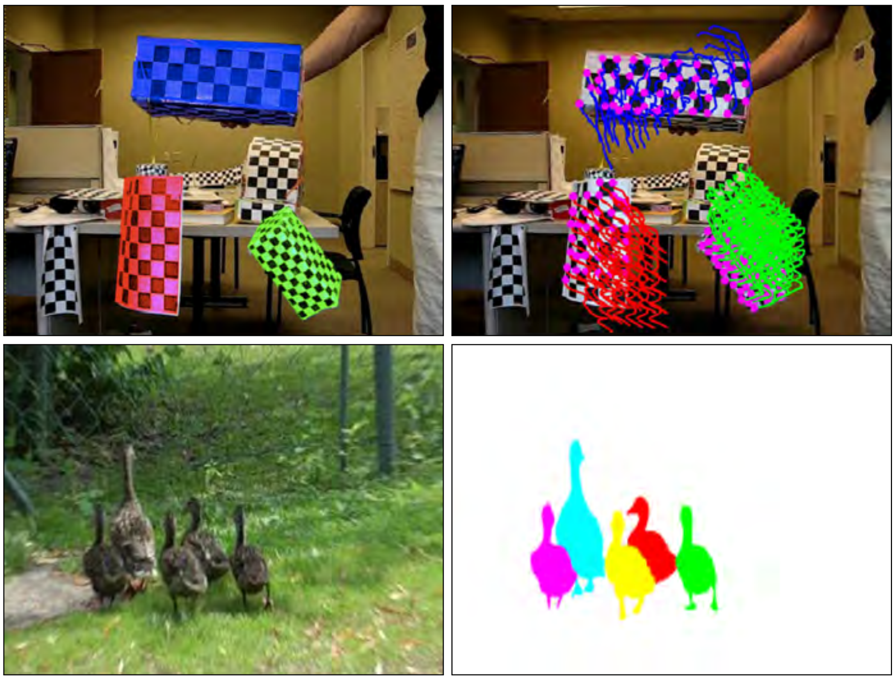


Figure 2.1: **Top Row:** Trajectory-based dataset sequence [38]. The original image is on the left, whereas the trajectory-based motion image is on the right. **Bottom Row:** Region-based dataset sequence [39]. The original image is on the left and the region-based annotation of the moving objects on the right.

two distinct types: trajectory-based and region-based datasets. An example of both groups is shown in Fig. 2.1.

- **Trajectory-based datasets:** In trajectory-based datasets, moving objects are represented by a limited number of salient points features, already extracted and tracked throughout the video sequence. These trajectories represent only a part of the moving object, hence, the motion can be tracked even under partial occlusion. The trajectories belonging to each motion are grouped together with a unique label to perform motion annotation.

- **Region-based datasets:** On the contrary, the region-based datasets do not use sparse points but perform pixel-wise motion labeling. The result is a more precise annotated segmentation of moving object contours, but the occlusion problem becomes harder to solve. The main drawback of region-based datasets is the heavy computational time required to generate the label for each pixel.

Both trajectory- and region-based datasets have paved the way for valuable contributions in the motion segmentation research. Each presented dataset had its own

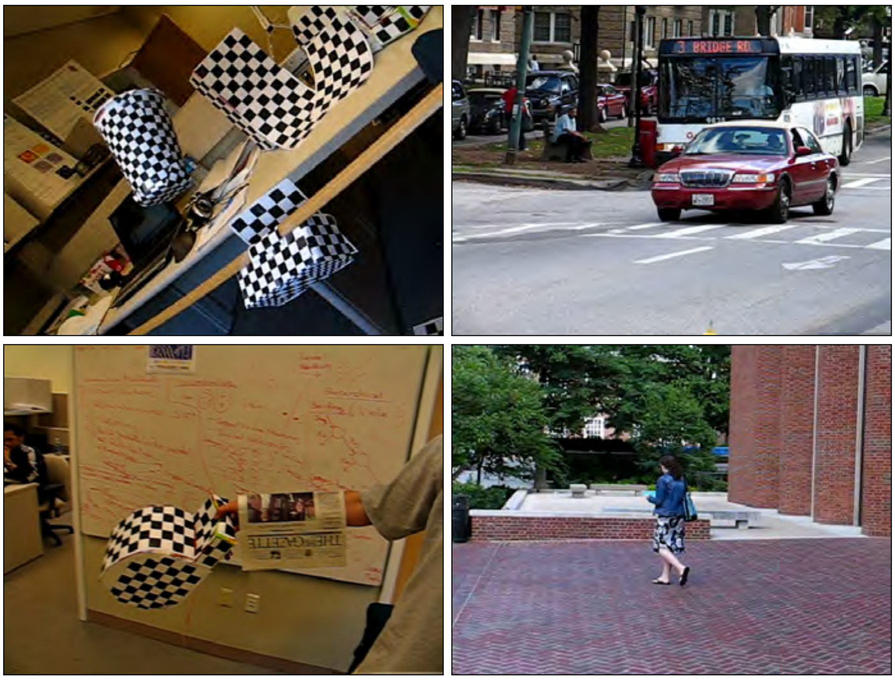


Figure 2.2: **Top Left:** A frame from a checkers board synthetic sequence. It acts as rigid motion. **Top Right:** A frame from an outdoor car real sequence. It contains two rigid motions. **Bottom Left:** A frame from a checkers board synthetic arm sequence. It contains articulate motion. **Bottom Right:** A frame from a people real sequence. It contains non-rigid motion.

strengths and limitations. A list of a few widely known datasets along with their features and properties is given in Table 2.1. The details of some of them are:

1. **Hopkins155 [38]:** Among both types, the most well-known publicly available dataset designed to address the motion segmentation problem was the trajectory-based Hopkins155 [38] dataset. It is still considered to be a reference benchmark for MS. This trajectory-based dataset was widely accepted in the community because it provided a simplified method for researchers to easily compare their algorithms. Hopkins155 comprised 155 short sequences having 2 or 3 motions, as illustrated in Table 2.1. More than 100 video shots were synthetic i.e. having checkerboard and books passively moved by humans. The remaining sequences included rigid motions of moving traffic, non-rigid motion of people and articulated motion of cranes, human arm and head. A few examples of synthetic as well as real images of Hopkins155 are given in Fig. 2.2. All the trajectories in all the sequences were complete, meaning that the moving objects do not get occluded, neither do they go out of the field of view. The ground-truth was provided as a motion or background label for all the trajectories.

The characteristics of Hopkins155 had a significant impact on MS research. All the algorithms were evaluated on a single dataset with a robust metric [38]. But the inherent limitations in Hopkins restrained the development of MS algorithms further. The dataset contained no missing data, the length of videos was extremely short and the number of motions captured per sequence were few. The dataset comprised predominantly synthetic video shots, and the kind of real noise like occlusion, stopping motion, perspective distortion, multiple appearance-disappearance of objects etc, that regularly occur in real-life videos was not present. In 'Hopkins Additional', which was a name given to the 16 sequences with missing data added to Hopkins155, an effort was made to overcome these limitations, but even in these sequences, 12 were synthetic. The number of motions were few, and the length of the sequences was still short. Misclassification levels reached on Hopkins nowadays are as low as 0.8% [31], which means that there is no room left to quantitatively distinguish one algorithm from the other. Therefore, although Hopkins paved the way to meaningfully analyze MS algorithms, in terms of modern day requirements, the dataset has become obsolete.

2. **FBMS59 [39]:** To overcome the innate limitations present in Hopkins155, a region-based motion segmentation dataset **BMS26** [50], comprising 26 sequences was presented. This dataset mostly contained snippets of movies with 1 to 2 people as moving objects captured from predominantly a moving camera. Some sequences were also borrowed from Hopkins155. Later, 33 more video shots were added to this dataset to create FBMS59 [39], having 59 sequences in total. Every 20th frame in FBMS59 came with a pixel-accurate ground-truth segmentation of moving objects. Thus, the ground-truth was available only on approximately 5% of the complete dataset. Some salient features of the dataset can be seen in Table 2.1. The dataset encapsulated rigid motion of cars, non-rigid and articulated motion of people and animals. Some video shots were captured at varying light conditions, and most sequences contained partial occlusion. A few sample frames of the FBMS59 dataset are given in Fig. 2.3.

   Although *FBMS59* tried to address the constraints in Hopkins, the fact that the ground-truth was available only on 5% of the whole dataset makes its usage restricted. There was an increase in sequence length but the average number of motions per sequence was still low. The problem of real noise was also addressed partially as there were sequences with partial occlusion, but rarely any with complete occlusion. The effect of perspective distortion was vaguely captured, and

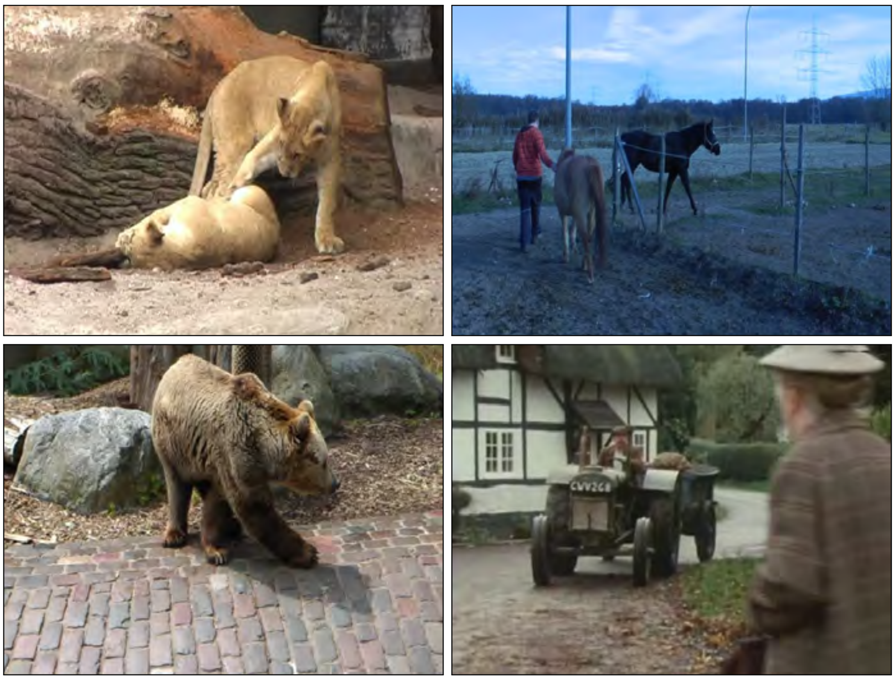


Figure 2.3: **Top Left:** A frame from the lions sequence. It contains two non-rigid motions. **Top Right:** A frame from the horses sequence. It contains three non-rigid motions. **Bottom Left:** A frame from the bears sequence. It contains a single non-rigid motions. **Bottom Right:** A frame from the marples sequence. It contains two motions, one rigid and one non-rigid.

the aspect of stopping motion and multiple appearance-disappearance of objects were not tackled.

3. **MOViCS [51]**: In this dataset 11 short sequences, with non-rigid real motions, were presented with minimal motion complexity. A region-based annotation was provided on a few frames, as the time cost of annotating all the frames would have been enormous. The ground-truth was provided only on approximately 29% of the complete dataset. Some salient features of the dataset can be seen in Table 2.1. The dataset encapsulated non-rigid and articulated motion of animals. Some video shots were captured with slight variation in light conditions, and most sequences contained partial self-occlusion. Some examples of the images present in this dataset are given in Fig. 2.4.

   The MOViCS dataset contributes to the already existing datasets by adding a few sequences containing non-rigid motions, but in itself, it cannot be treated as a comprehensive dataset. The length of the sequences is quite small, with an average 47 frames per sequence. The average number of motions is only 2.

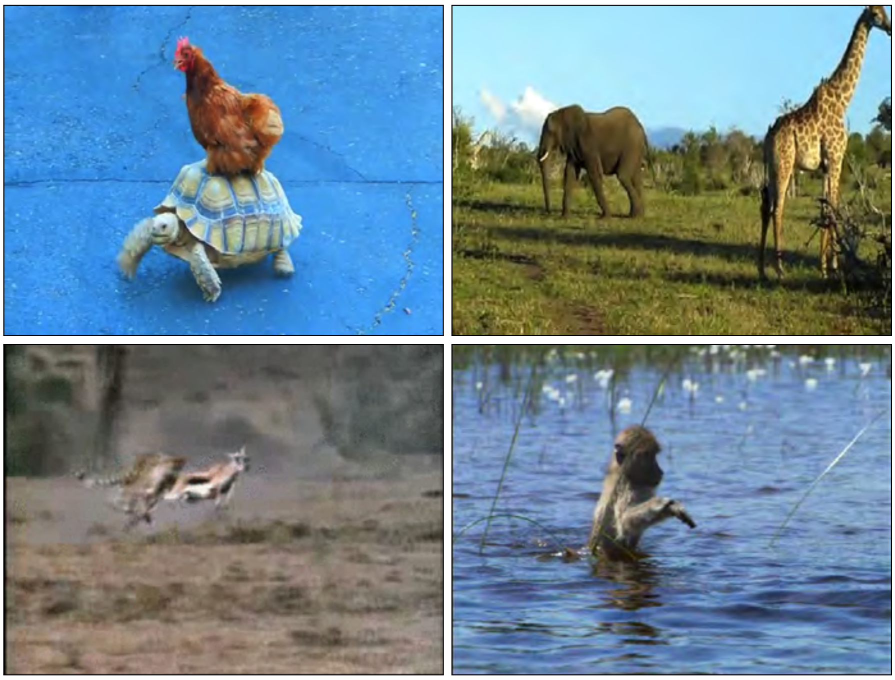


Figure 2.4: **Top Left:** A frame from the chicken-on-turtle sequence of MOViCS dataset. It contains two non-rigid motions. **Top Right:** A frame from the elephant_giraffe_all2 sequence of MOViCS dataset. It contains two non-rigid motions. **Bottom Left:** A frame from the cheetah sequence of SegTrack dataset. It contains two non-rigid motions. **Bottom Right:** A frame from the monkey sequence of SegTrack dataset. It contains one non-rigid articulate motion.

Hence, although MOViCS brings forth some non-rigid motion sequences to test algorithms performance on them, it lacks the capability to rigorously test all the features desired in a MS methodology.

4. **SegTrack [41, 52]**: In this dataset, 13 short sequences, with non-rigid real motions, were presented with limited motion complexity. In this dataset as well, a region-based annotation was provided but in contrast with FBMS59 and MOViCS, the grud-truth was provided on all frames. Some salient features of the dataset can be seen in Table 2.1. The dataset encapsulated non-rigid and articulated motion of people and animals. One sequence contained the motion of a parachute. Generally, video shots were captured with slight variation in light conditions, and most sequences contained partial occlusion. A few sample frames of SegTrack dataset are shown in Fig. 2.4.

   The SegTrack dataset also has limited contribution to be considerd as a comprehensive MS dataset. Though, the provided ground-truth on all frames makes its usage quantitatively significant. The length of the sequences is small, with an average 76 frames per sequence. Like MOViCS, the overall average number of motions per sequence is also only 2. SegTrack lacks sequences with rigid-motions. It also does not contain motions undergoing complete occlusion, stopping motion and multiple-appearance-disappearance. Hence, it too lacks the depth in sequence variations and complexity to rigorously test the comprehensiveness of a MS algorithm.

5. **Other Datasets**: There are other datasets, which were created for Video Segmentation or related fields, and were also used for MS research. These datasets include **VSB100 [53, 54]**, **Extended-Yale benchmark** [**40**] and **Cambridge CamShift dataset** [**42**]. All these datasets have limited number of motions or frames per sequence, which makes their use in MS community quite limited. The annotation methodology in all datasets changes with reference to the motion complexity, which determines the processing purpose.

It is apparent that researchers have focused a lot on MS datasets. Many proposals with variable features and properties were proposed. Baring Hopkins155 and FBMS59, others provide supplementary contribution to the motion segmentation problem. Hopking155 has been exhaustively tested and has lately become obsolete, as it lacks the capability to further evaluate algorithm's improvement. Although some limitations of Hopkins155 were addressed in FBMS59, there is still a requirement for a comprehensive dataset, which can tackle the prevalent constraints of the two datasets.

2.2 Motion segmentation methodologies

Motion segmentation is a preprocessing step for many computer vision tasks. Owing to this feature, it has remained of particular interest for researchers. In recent years, many methodologies have been put forth [25–29]. The problem has been addressed in a variety of approaches. A category-wise explanation of some of the more reputed methodologies is as follows,

- **Subspace clustering:** MS problem was initially approached as a subspace affinity problem in [30]. This approach was then extended as a sparse subspace clustering (SSC) problem [31] by proposing the use of sparse representation of data drawn from multiple low-dimensional linear or affine subspaces. SSC reached as low as 0.8% misclassification on Hopkins155. Similarly, a low-rank representation (LRR) [32,33] to segment data drawn from a union of multiple linear or affine subspaces, was proposed. Unlike the sparse representation (SSC), which computes

Table 2.1: A summary of the features and properties of the state-of-the-art MS datasets. Acronyms are **Avg.:** Average, **Max.:** Maximum, all averages and maximums are per sequence. In Occlusion, **P:** Partial, **C:** Complete, **M:** Multiple, **M.A.D.:** Multiple Appearance-Disappearance of objects.

State-of-the-art MS Datasets													
Datasets	**Dataset Features**							**Dataset Properties**					
	Total sequences	Total frames	Avg. frames	Total motions	Avg. no. of motions	Max. no. of motions	Ground-Truth (%)	Real objects (%)	Occlusion (P/C/M)	Missing data	Perspective distortion	Stopping motion	M.A.D. of objects
Hopkins[38]	155	4615	29.8	345	2.2	3	100%	33%	**x**	**x**	**x**	**x**	**x**
FBMS[39]	59	13860	234.9	193	3.3	7	5%	100%	**P**	✓	✓	**x**	**x**
SegTrack[41, 52]	13	992	76.3	22	1.7	6	100%	100%	**P**	✓	**x**	**x**	**x**
MOViCS [51]	11	514	46.7	17	1.5	2	20%	100%	**P**	✓	**x**	**x**	**x**

the individual sparsest representation of each data vector, LRR aimed to find the lowest-rank representation of a collection of vectors jointly. LRR reached as low as 3% misclassification on Hopkins155. Another method with a new interpretation to extract the rank of trajectory matrix and an improved affinity measure was proposed [34–36] known as Enhanced Local Subspace Affinity (ELSA) and Automatic-Adaptive subspace Affinity (A-ASA). ELSA and A-ASA reached results as low as 0.96% on Hopkins155. The approach proposed in [37] introduced a temporal subspace clustering method for unsupervised segmentation of human motion by designing a temporal Laplacian regularization function to model the sequential information in time. This method had limited applicability in general MS problems.

- **Handling missing data:** The issue of corrupted trajectories and missing data was addressed in a mathematical framework as Agglomerative Lossy Compression (ALC) [43]. This algorithm consists of minimizing a cost function by grouping together trajectories. The cost function is given by the amount of information required to represent each manifold, summed over all of the manifolds. ALC reached a minimum misclassification of 3.37% on Hopkins155. This technique does not guarantee to find the global maximum. Another problem is the need to tune a parameter, which depends on the noise level of the input sequence and on the number of clusters of the video. Besides ALC, another approach presented an algebraic geometric approach as Generalized Principal Component Analysis (GPCA) [55]. This technique used GPCA to fit a polynomial of degree N, where N is the number of subspaces (i.e. the number of motions), through the data and estimate the bases of the subspaces using the derivatives of the polynomial. This algorithm was also able to deal with missing data.

- **Statistical and layer techniques:** MS was also approached by statistical techniques as well [56], where a framework for two-view segmentation was presented. A local sampling based hypothesis for the estimation of fundamental matrices was generated. Using the hypothesis, a combinatorial model is created and then optimized. The authors comment that the outcome of the method heavily depends on the initial set of hypotheses. In addition to statistical methods, MS was also approached through layers techniques [57–59]. Specifically, [57] proposed an approach to extract motion layers from a pair of images with large disparity motion. A topological clustering algorithm along with Scale-Invariant Feature Transform (SIFT features) [60] establishes motion models. Using an affine transformations model for each cluster, a graph-cut based algorithm was employed to segment the scene into several motion layers. This technique had a problem when handling occlusions and was sensitive to change in the illumination of the scene. Some approaches dealt with moving object segmentation as an energy minimization graph problem. In [61], a fully connected spatio-temporal graph was built over object proposals. Expectation maximization of the energy term that incorporated confident sparse long-range feature tracks is done, to ensure similar temporal labeling of objects. This approach is dependent on image resolution. On the contrary, in [62], the problem was posed as a minimum cost multicut formulation for motion trajectory segmentation. The costs were defined as positive or negative edge weights.

- **Tracking and Segmentation:** Recently, two approaches proposed joint tracking and segmentation of moving objects in videos. One presented an algorithm which integrates the multi-part, tracking and segmentation into a unified energy optimization framework [63].This approach is able to deal with a single object per frame and per video. The other approach, proposed a multi-target tracker that exploited low level image information and associated each super-pixel to a specific target object or background [64]. It needs annotated silhouettes masks to obtain object specific foreground for segmentation.

- **Optical flow:** Moreover, optical flow based long term analysis of point trajectories was also performed for moving object segmentation [50,65], and to turn these point trajectories into dense motion regions a hierarchical variational approach was introduced recently [39, 66]. Another optical flow based approach [67] presented multiple figure-ground segmentations on motion boundaries and ranked them based on a Moving Objectness Detector, with the final segmentation as the top ranked spatio-temporal tube. Another proposal used optical flow to infer long-term occlusion relations from video, and framed it as a convex optimization problem to segment image domains [25].The work presented in [68] proposed to detect disoccluded regions by inferring piecewise smooth deformation fields. By using motion and appearance cues, it partitioned the disoccluded region and grouped its components with the object.

The recent inclination of the community is towards optical flow based methods because the problem is then dealt in totality for all motions and for all distortions, especially in the case of missing data. The computational time taken for optical flow based methods remained a limitation for a long time, but the rise in high speed graphics processing unit (GPU) processors containing multiple cores has made it possible for these methods to be considered for usage in real time.

2.3 Motion annotation

2.3.1 Motion annotation approaches

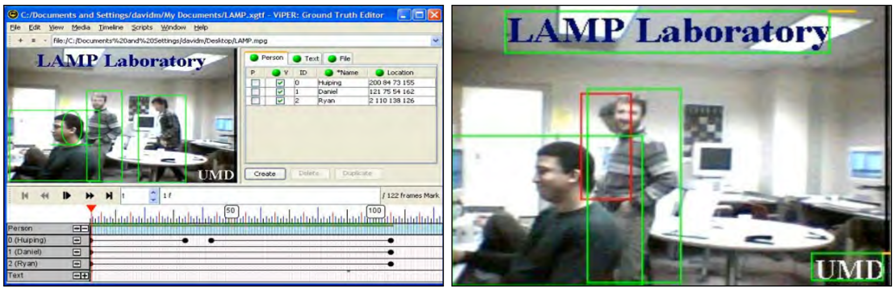
In general, the solutions of video annotation problem try to achieve two distinct objectives, either to reduce the expert-user annotation time in generating the ground-truth of large scale video data, or to improve annotation quality, or both. These objectives are achieved by two distinct approaches. One approach is to put forth comprehensive video annotation platform tool as a standalone package, which can label motions or objects of interest in video sequences. The other proposition is to devise label propagation methodologies, which can be incorporated in existing tools. The state-of-the-art in video annotation includes techniques from both practices.

2.3.2 Standalone annotation packages

Regarding standalone packages, several video annotation tools have been developed in recent years. Predominantly computer vision and machine learning methods are used as support for efficient human annotation. The different tools can be distinguished based on the functionalities they support. Some of the more used packages are listed below,

- **ViPER:** The pioneering work on video annotation was presented in ViPER [69, 70], which was a reconfigurable video performance evaluation resource. It provided an interface for manual ground-truth generation, an evaluation metric and a visualization tool as shown in Fig. 2.5. It was a Java based desktop application, which propagated rectangular or polygon region-of-interest (ROI) through linear interpolation.

- **GTTOOL and GTTOOL-W:** Two similar desktop-based GTTOOL [71] and web-based GTTOOL-W [72] tools were presented, with a goal to improve user experience with respect to ViPER [69] by providing edit shortcuts, and by integrating some basic computer vision algorithms to automate. The collaborative web-based implementation featured an easy and intuitive user interface that



C:/Documents and Settings/davidm/My Documents/LAMP.xgtf - ViPER: Ground Truth Editor

file:/C:/Documents%20and%20Settings/davidm/Desktop/LAMP.mpg

P	V	ID	*Name	Location
□	☑	0	Huiping	200 84 73 155
□	☑	1	Daniel	121 75 54 162
□	☑	2	Ryan	2 110 138 126

Figure 2.5: **Left:** The GUI of the ViPER annotation tool, with the video canvas on the left, the spreadsheet on the top right and the remote is in the middle. **Right:** The spatial annotations editing on the video canvas [69, 70].

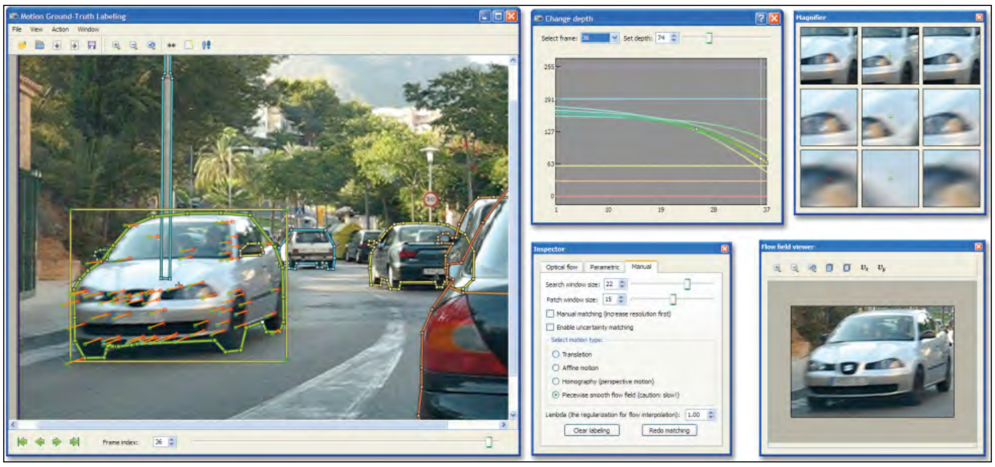
allowed instant sharing/integration of the generated ground-truths. The label propagation in these tools were performed using tracking approaches.

- **LableMe-Video:** Relatively recently, a popular online, openly accessible tool LableMe-Video [73], was presented that allows annotation of object category, motion and activity information in real-world videos. This tool used homography to propagate the label across key frames in the video. Using this system, a scalable video database composed of diverse video samples and paired with human-guided annotations was built. It lacks uniformity in annotation, as a user with out any qualification on annotation expertise, can annotate to add to the dataset.

- **iVAT:** With the same focus, iVAT [74], an interactive Video Annotation Tool, which supports manual, semi-automatic and automatic annotations was presented. This tool integrated several computer vision algorithms working in an interactive and incremental learning framework. This makes the tool flexible and suitable to be used in different application domains. The main limitation of the tool is in annotating long term motions with partial and complete occlusions. It works well for small motions.

- **HIL:** Another human-in-loop (HIL) methodology [45], to create ground-truth for videos containing both indoor and outdoor scenes was put forth with the idea that human beings are experts at segmenting objects and inspecting the match between two frames. The approach contained an interactive computer vision system to allow a user to efficiently annotate motion. The GUI with other utilities is shown in Fig. 2.6.

The standalone packages provide a single platform for moving object or object of interest annotation. They possess the inherent trait of providing all the annotation



Select frame:
Set depth: 74

Search window size: 22

Patch window size: 15

□ Manual matching (increase resolution first)

□ Enable uncertainty matching

Select motion type:

○ Translation

○ Affine motion

○ Homography (perspective motion)

● Piecewise smooth flow field (caution: slow!)

Lambda (the regularization for flow interpolation): 1.00

Frame index: 36

Figure 2.6: A screen shot of the human-in-loop motion annotation system. The main window, depth controller, magnifier, flow field viewer and control panel [45].

utilities on a user-friendly Graphical User Interface (GUI). A general limitation of such platforms is the flexibility to accommodate further algorithms, which can enhance their annotation range. This limitation makes them a time-bound contribution, as with the passage of time their utilities become obsolete with respect to the modern requirements of annotation.

2.3.3 Label propagation techniques

The other approach as mentioned in Section 2.3.1 are label propagation methodologies. This is a relatively new track to solve the annotation problem [75–79]. Some recent work has been presented related to label propagation, where a manually given object label in key frames is propagated forward and/or backward in all the frames the object exists. This problem was also tackled in several different ways as well.

- **Probabilistic:** Probabilistic graphical models for propagating labels in video sequences were used in [75]. An Expectation Maximization (EM) algorithm propagates the labels in a chunk of video with start and end frames already labeled. The unlabeled parts of the video are dealt with in a batch setting. In [76], a similar approach was used to train a multi-class classifier. The pixel labels estimated by the trained classifier were fed into a Bayesian network for a definitive iteration of label inference. A hybrid of generative propagation and discriminative classification in a pseudo time-symmetric video model enables conservative occlusion handling. Moreover, in [77] the limitations of pure motion and appearance based

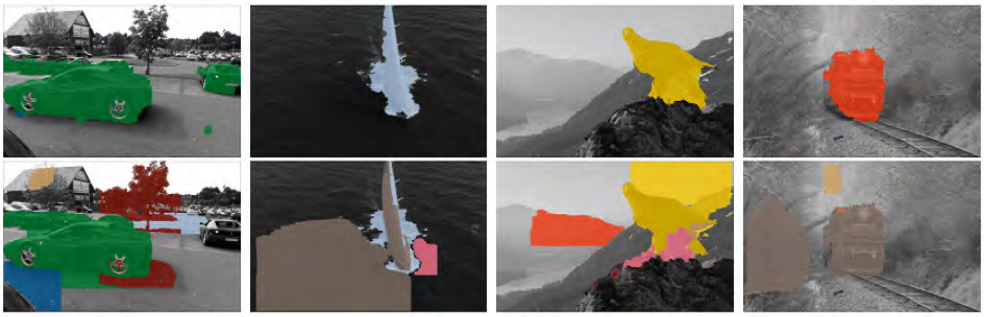


Figure 2.7: **Top:** Sample frames with our annotation using 20% of the sequence. **Bottom:** Probabilistic baseline labeling on the same frames. Different colors correspond to different object classes [79].

propagation methods were shown, especially the fact that their performances vary on different type of videos. To avoid these limitations, a probabilistic framework was proposed that estimated the reliability of the appearance-based or optical flow-based label sources and automatically adjusted the weights between them.

- **Active Frame Selection:** An active frame selection approach was adapted in [78, 79]. In [78], active frame selection is done by selecting k frames for manual labeling such that automatic pixel-level label propagation can proceed with minimal expected error. Here the frame selection criterion is joined with the predicted errors of a flow-based random field propagation model. The method excels in utilizing human time for video labeling effectively. In contrast, an information-driven active frame, location and detector selection approach was used in [79]. The method optimizes on a given uncertainty bound, the selection of a detector at a particular location and also minimizes label uncertainty at each pixel. Its results in comparison with a probabilistic baseline are shown in Fig. 2.7, as taken from their original paper [79]. It also tries to optimize for computational cost for both manual and semi-automatic labeling.

- **Graph Methods:** More recently, a semi-supervised video annotation approach [80] was proposed by learning an optimal graph from partially labeled object. The methodology also exploited multiple features, which could accurately embed the relationships among the data points. The similarity graph used the geometrical relationships among the training data points. The model was extended to address out-of-sample and noisy label issues. From another perspective, a diffusion approach for label propagation was used in [81]. The application of anisotropic diffusion on graphs and the corresponding label propagation algorithm, on the vector bundles of Riemannian manifolds was presented. This definition of new diffusivity operators significantly improved semi-supervised learning performance.

- **Inferring Color:** For an application of color label propagation, in [82], the problem of inferring color composition of the intrinsic reflectance of objects was addressed. The color labels were propagated between regions sharing the same reflectance, and the direction of propagation was propagated to be from regions under full illumination and normal view angles to abnormal regions.

The literature on label propagation has had some new developments but there are still challenges, which are still unsolved. Most of the state-of-the-art techniques revolve around taking key frames and propagating the labels, but they are rarely able to deal with distortions in motion.

2.4 Conclusion

The existing datasets, due to their limitations, cannot perform an analysis of algorithms in the presence of multiple occlusions, stopping motion, perspective distortion, multiple appearance-disappearance, and real-life noise of camera motion. Moreover, if a new difficult benchmark is introduced with long sequences, then most of the state-of-the-art algorithms will not be applicable on it. Hence, there is a vaccum, which can be filled by creating a comprehensive new benchmark dataset, focusing on both long and short sequences with real objects, increased number of motions per sequence, and real distortions. In this way, prevailing state-of-the-art algorithms can be applied and analyzed on short sequences with real noise, while new algorithms can be designed and then tested on long sequences. Keeping this in view, a major contribution is made in filling this vacuum with the work presented in Chapter. 3.

The existing methodologies in label propagation address the problem in a limited range of applications. Though, they perform well, they lack utility in real life long videos in outdoor scenes, where multiple occlusions, stopping motion, perspective distortion, multiple appearance-disappearance and noise of camera motion, are present. A reason for these limitations is the absence of a video dataset, where these optical phenomenons could be tested. With the benchmark presented in chapter. 3, it is possible to test these annotation limitations quantitatively. Results in the state-of-the-art demonstrate that the use of the semi-automatic, as well as the automatic, modality in annotation drastically reduces the human effort while preserving the quality of the annotations. Related to this viewpoint, two domains are addressed in chapter. 4. In the first part, a comprehensive framework for trajectory-based motion annotation tool is presented. In the second part, a motion-region annotation technique is presented, which complements the HIL annotation tool [45]. Our motion region annotation methodology is generally applicable on objects undergoing partial occlusion by static occluders, with a limited application on objects undergoing occlusion by other moving objects. In chapter. 4 a consolidated evaluation is also performed to establish the usage of the scheme in real life scenes.

In God we trust; all others must bring data.

William Edwards Deming

Chapter 3

A collection of challenging motion segmentation benchmark datasets

All the MS datasets presented in the literature review of Chapter 2 are not without limitations. The boundaries of motion segmentation algorithms can only be pushed if these pointed out limitations are addressed. Both, the trajectory- and region-based datasets, lack in sequences with long term motion undergoing partial, complete or multiple occlusion. Occlusions of these types regularly occur in daily life scenes. Datasets with limited inclusion of this phenomenon do not possess the capability to test the algorithms in this domain. The same limitation is present for other distortions as well, which include illumination changes, shadows, stopping motion, multiple-appearance-disappearance, perspective distortion, drastic depth change and distortion due to camera motion, jitter or others. These limitations along with the idea to provide complete ground-truth on all the motions on all the frames inculcated the motivation to build a new benchmark dataset.

Our aim was set to build two trajectory- and region-based benchmark datasets, each comprising of long and short sequences. In this chapter, first the description of the acquisition setup to build the dataset is given. After that, the trajectory-based long and short sequences database is explained, followed by the presentation of the region-based long and short sequences database. The features and properties of each dataset are listed in Tables 3.1 and 3.2, respectively. Afterwards, the benchmarking setup and the methods used to perform this first quantitative evaluation are presented, including a discussion on both quantitative and qualitative performance of the analyzed state-

Table 3.1: A summary of the features of our trajectory-based and region-based benchmark MS datasets. The details of both long and short sequences are listed. Acronyms are **Avg.:** Average, **Max.:** Maximum, all averages and maximums are per sequence

Proposed Trajectory-based, and Region-based MS Datasets							
Datasets	**Dataset Features**						
	Total sequences	Total frames	Avg. frames	Total motions	Avg. no. of motions	Max. no. of motions	Ground-Truth (%)
Trajectory-based Long	19	15538	817.8	201	10.6	17	100%
Trajectory-based Short	162	6942	42.9	442	2.7	6	100%
Region-based Long	20	16300	815.0	235	11.8	23	100%
Region-based Short	150	6262	41.7	440	2.9	6	100%

of-the-art methods. In the end, the conclusion includes a summary of the presented benchmark with a comment on its usage and current limitations.

3.1 Data acquisition

The first step in building any video dataset is to capture the video shots through a consistent reference setup. To build our database, the acquisition of all the sequences was performed in high definition (HD) at 1920x1080 pixels per frame at 30fps. The processing of long HD videos is computationally extremely expensive, especially in methods involving optical flow. Therefore, although all the captured sequences are provided in original resolution with the database, for processing purpose the frame size is reduced to 640x480. This reduction in resolution does not effect the performance analysis of any algorithm. Any methodology, which works well on 640x480 will technically perform the same on HD videos, except for the computational time. Even the ground-truth on the low resolution videos can easily be linearly interpolated on HD video, while remaining inside a minimum error bound.

All the sequences are video shots of real-life natural scenes which we come across

Table 3.2: A summary of the properties of our trajectory-based and region-based benchmark MS datasets. The details of both long and short sequences are listed. Acronyms are, in Occlusion, **P:** Partial, **C:** Complete, in Weather, **S:** Sunny, **Cl:** Cloudy, **D:** Dark, **M.A.D.:** Multiple-Appearance-Disappearance of objects

Proposed Trajectory-based, and Region-based MS Datasets									
Datasets	**Dataset Properties**								
	Real objects (%)	Occlusion (P/C/M)	Weather (S/Cl/D)	Missing data	Perspective distortion	Stopping motion	Shadows	Illumination Change	M.A.D. of objects
Trajectory-based Long	100%	P/C/M	S/Cl/D	✓	✓	✓	✓	✓	✓
Trajectory-based Short	100%	P/M	S/Cl	✓	✓	✓	✓	x	x
Region-based Long	100%	P/C/M	S/Cl/D	✓	✓	✓	✓	✓	✓
Region-based Short	100%	P/M	S/Cl	✓	✓	✓	✓	x	x

in our daily lives. The moving objects in the videos are cars, motorbikes, bicycles and people. In the database, there is a diverse collection of rigid and non-rigid motions. From a total of 26 video shots, 19 sequences contain trajectory-based ground-truth, and 20 contain region-based ground-truth. Hence, 13 sequences contain both trajectory- and region-based ground truth. Among all, 22 video shots were captured while standing or walking, and 4 were captured with the camera inside a moving car. The sequences captured from inside a moving car introduced a relative motion between the camera and the outdoor scene. Two sequences, one in each benchmark, were captured as tower camera view of the scene. The videos are of variable length ranging from 460 frames (15 sec) to 1737 frames (58 sec) at 30fps. The number of captured motions per sequence, excluding the camera motion, are minimum 4 and maximum 23. The maximum number of motions per frame reaches up to 6 in some sequences.

All sequences contain real moving objects, while their motion possesses a combination of distortions including occlusion, variable weather conditions, missing data, perspective distortion, stopping motion, shadows, illumination changes and multiple appearance disappearance of objects.

- Occlusion: Moving objects while in motion get partially or completely occluded by other objects, static and/or moving. These situations create partial, complete or multiple occlusion of a single moving object.

- Weather: The acquisition of sequences is performed in variable weather conditions, i.e. in sunny weather, in cloudy weather and in the evening, when the sun is about to set, giving rise to a darkish backdrop.

- Missing data: When trajectories, or areas in motion regions, disappear along the motion of an object due to occlusion, tracking failure or other distortions, it results in missing data.

- Perspective distortion: It is captured in the sequences, where the distance of the objects from the camera changes considerably with respect to the angle of view of the camera, while keeping the object inside the field of view (FoV) throughout.

- Stopping motion: When a moving object, instead of being in continuous motion throughout its appearance in the sequence, remains static in some frames of the sequence, this motion results in stopping motion.

- Shadows: Motion of objects, with the sun on top or behind the object, casts their shadow between the object and the camera. These shadows are captured in several sequences.

- Illumination Changes: The effect of illumination change is captured in multiple sequences, either due to the slight change in lighting conditions during object motion or due to the considerable change in depth of the moving object during the course of its motion.

- Multiple Appearance-Disappearance: These are the instances, when a moving object while in motion goes out of the FoV, and comes back in again in the sequence, giving rise to appearing, disappearing and then re-appearing of the object.

The presence of these real-life distortions make the database diverse and challenging. Any algorithm tested on this proposed dataset needs to encapsulate multiple distortion handling capabilities to be able to exhibit good results. Hence, the performance measures of less than 1%, which became a norm for Hopkins155, are not expected. Instead, the difficulty would push the boundaries of MS methodologies diversity.

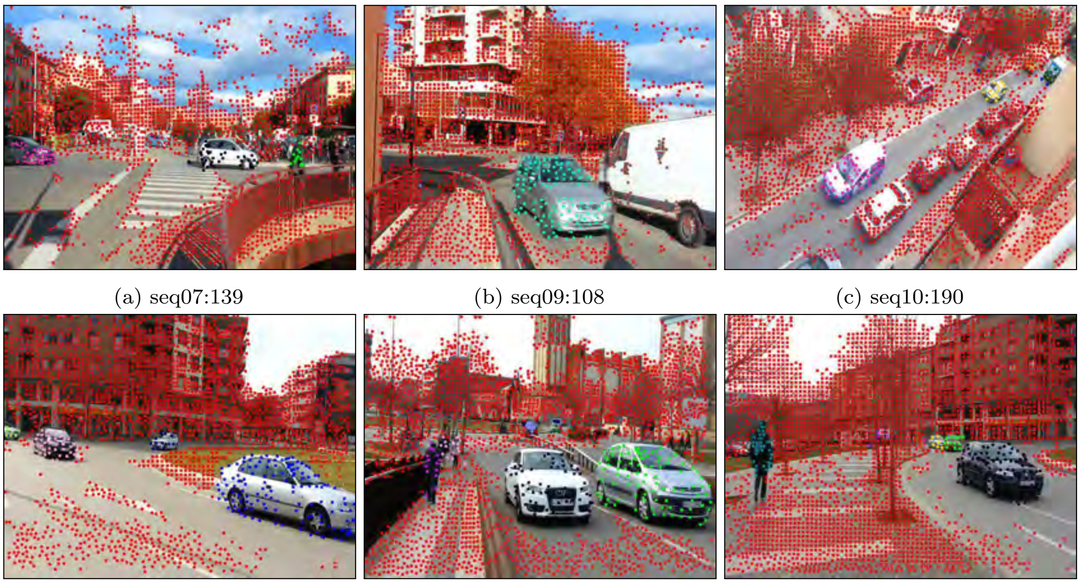
3.2 Trajectory-based datasets

The proposed trajectory-based datasets in their composition are similar to Hopkins155. It primarily contains 19 long sequences in which sparse point trajectories on all the motions are tracked and then labeled to form the dataset. These long sequences with ground-truth are further cut into 162 small sequences, so that they are processable by the current algorithms.

3.2.1 Trajectory-based long sequences

The dataset is formed, firstly, by the acquisition of 19 long real-life natural scene sequences. After acquisition, two more steps are performed, i.e. Tracking, and Annotation. Tracking is necessary to form a trajectory-based dataset, as it is used to capture the motion of the moving objects in the form of sparse trajectories. The camera motion is also captured by having trajectories on the background. These trajectories are then annotated to form the ground-truth of the dataset.

Tracking: A tracker should be able to extract robust and consistent trajectories. Its choice is dependent on the type of motions to be captured and the kind of distortions



(a) seq07:139 (b) seq09:108 (c) seq10:190

(d) seq13:478 (e) seq14:693 (f) seq15:30

Figure 3.1: Sample frames of 6 sequences in trajectory-based long sequences with ground-truth overlay. In the sub-figure description code 'seqXX:YYY', XX is the sequence number and YYY is the frame number.

these motions contain. The acquired sequences include diverse motion types, comprising small and large rigid and non-rigid objects (e.g. cars, bikes, people individually or in groups). The distortions include partial/complete/multiple occlusion, illumination changes, shadows, perspective distortion, stopping motion, and multiple appearance-disappearance. It is an extremely difficult task to robustly capture these characteristics, especially with a uniform distribution of trajectories over object regions.

Many recent tracking approaches exhibited relatively promising results. In [83], an optical flow based tracking algorithm is presented, which enables robust extraction of densely sampled trajectories. These trajectories are converted into descriptors for action recognition by limiting the length for each trajectory to 15 frames. If used for longer frame lengths, the algorithm results in undesirable sparse trajectories on homogeneous regions of moving objects.

Recently, a promising Large Displacement Optical Flow (LDOF) based tracking algorithm was proposed [46] with a publicly available GPU-accelerated implementation [84]. The algorithm tracks robust point trajectories over densely sampled regions using LDOF. The trajectories on rigid and non-rigid objects are consistent. It stops sampling in homogeneous regions of background, while retaining densely sampled trajectories on homogeneous regions of moving objects. We used this algorithm for tracking in our dataset. The property of this algorithm of providing dense trajectories in moving object regions was of particular use for our trajectory-based dataset. The robust tracking of long term trajectories and re-sampling in case of occlusion, made sure that there were densely populated trajectories on all the moving objects in all frames during the course of its motion.

The computational cost of LDOF calculation is high but once LDOF is known, dense point sampling and tracking is fast. Tracking can be carried out for samples of variable size, depending on the required density of image coverage. When performing the initial tests tests on our sequences we observed that, if the sample size is kept high i.e. 16 or more, then there are less number of trajectories per moving object and small objects are missed. If the sample size is kept small i.e. 4 or less, then the total number of trajectories and the overall computation cost increases. Hence, we chose a sample size of 8 in our dataset as a trade-off between overall computational cost and dense coverage of small objects. We used *'mosegOB'* tracker executable [65] as it is a publicly shared implementation by the authors themselves. Its output is a post processed result in which outlying bad trajectories are filtered. The tracker output results in almost 0.39 million trajectories, capturing all the desired motions in all the sequences.

Some tracking results with ground-truth overlay can be seen in Fig. 3.1. In all the images, the motion of the rigid non-degenerate moving objects are captured with long trajectories having dense coverage. In Figs. 3.1a, 3.1e and 3.1f, trajectory labels on non-rigid degenerate moving objects can be seen. These trajectories are relatively sparse as compared to rigid non-degenerate objects, but representative nonetheless. There is also a subset of noisy trajectories, outliers, that capture motion of more than one moving object in the sequence. These outliers can potentially force classification errors in MS algorithms because of the inherent multiple motions in them. Consistent shadows of moving objects are also tracked with the object.

Annotation. Annotating 0.39 million trajectories is a cumbersome task. This annotation problem is a topic of research called label propagation, in which manually assigned region labels in a few frames are propagated to the rest of the video. It is a challenging task, regarding which some recent work was presented in [75] and [76], as mentioned in Section 2.3 in Chapter 2. In [78], an active frame selection method is proposed to best utilize human effort by minimizing an error cost for any set of k manually labeled frames. However, if used on our dataset then multiple steps of pixel label propagation over multiple frames and then re-propagation of these region labels to trajectory labels result in unwanted accumulation of error.

We used a composite solution by first acquiring an approximation of segmentation labels on all the trajectories, and then performing manual correction to obtain an accurate ground-truth. The first step of approximation on all the trajectories is acquired by a recent MS algorithm [39, 65]. Afterwards, the wrongly estimated labels are manually corrected by a user with a ground-truth annotation software. The amount of time taken for manual effort is proportional to the number of wrongly segmented trajectories.

All tracks on a moving object are given a unique label. Two labeling principles are followed for all the sequences:

- Two separate objects with similar kind of motion are labeled uniquely if at any point in the sequence the object boundaries are visually separable.
- The noisy subset of trajectories that represent multiple motion in a single trajectory are assigned the label of the moving object they are most representative of.

The main features of our trajectory-based long sequences database are listed in Table 3.1, where it can be seen that 19 long sequences encompass more than 200 motions, spanning over 15500 frames, while containing all the real-life distortions. This contribution alone is more intricate than any MS dataset presented thus far.

3.2.2 Trajectory-based short sequences

Most MS algorithms were designed, since a long time, to be tested on Hopkins155 dataset. Due to this, most state-of-the-art algorithms only possess the capability to evaluate short sequences with no missing data. If the algorithms can not be tested on the long sequences dataset then the creation of a benchmark becomes a problem. This limitation of MS algorithms led us to the creation of a novel Hopkins155-like short sequences dataset. The length is shortened by cropping the long sequences into multiple short ones, making sure that in each of them real distortions are still prevalent.

In each short sequence, the trajectories as well as the ground-truth trajectory label are taken from its corresponding long sequence frames. All trajectories spanning 6 frames or less, are discarded as not representative enough. In total, we create 162 short sequences, which are generated from the 19 long sequences. The details of all the short sequences can be seen in Table 3.1, where we show that the 162 short sequences contain 442 motions, excluding background camera motion, spanning over 6900 frames. A separate subset of 40 sequences with no missing data is also created to process algorithms which are unable to handle missing data.

3.3 Region-based datasets

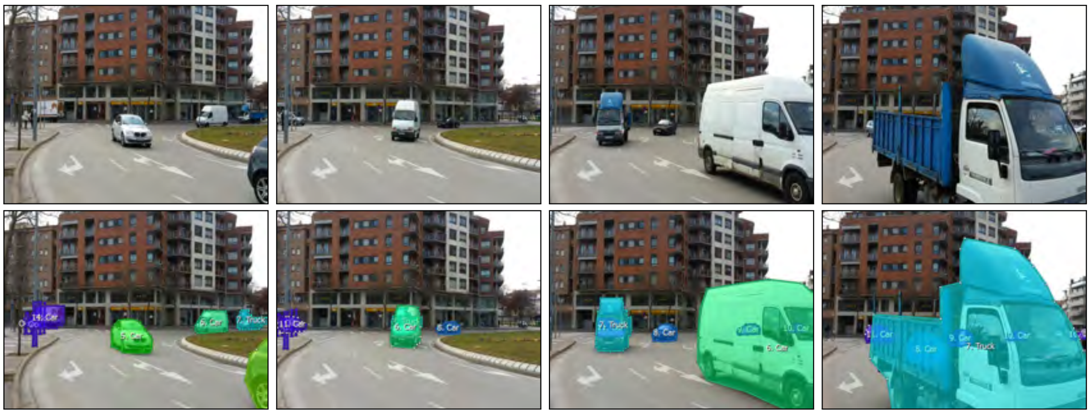
Over time, MS datasets have evolved from trajectory-based datasets [38] to region-based datasets [39, 41, 42, 52] as introduced in Chapter 2. In Hopkins155, instead of providing pixel motion labels on moving object regions in the sequence, motion labels on precomputed sparse trajectories were given. The purpose of trajectory-based dataset was to reduce the processing time, while still being able to capture motion. This approach added an extra tracking step besides acquisition and annotation. Due to the advancement in computational power, all the recently presented datasets contain pixel labels on moving object regions in every sequence. In our proposed region-based dataset, a total of 20 long sequences are presented with unique region label on all the motion regions of all the sequences. These sequences with ground-truth are further cut into 150 short sequences. An equivalent set of trajectories on every motion in both long and short sequences is also provided, so that they are processable for the state-of-the-art algorithms.

3.3.1 Region-based long sequences

The region-based dataset of long sequences is formed by first acquiring 20 long sequences that contain real distortions. In these sequences, the intermediate tracking step is not performed as trajectories are not needed. All the motion regions are directly labeled pixel-wise in each sequence through region annotation tool explained in the following section. Moreover, besides motion region labels, all the sequences contain trajectory-based ground-truth corresponding to each region label so that the state-of-the-art algorithms that only work on sparse trajectories can also be tested on this dataset.

Annotation It is an extremely difficult task to pixel-wise annotate all the motion regions in all the sequences. In recent works, researchers are trying to figure out methods with which a few manually annotated region labels in a few frames can be propagated to the rest of the frames, with some finite error bound [75, 76].

As mentioned in Section 3.2.1, in [78] an active frame selection method was proposed, where the authors minimize the error cost for any set of k manually labeled frames. This approach was not useful in our trajectory-based dataset due to multiple label transitions. Its application was problematic in our region-based dataset because it is an iterative method, which can become computationally expensive while searching for the best k frames. Its utility in sequences with less number of frames containing limited partial occlusion is present. It becomes increasingly unusable in our case because



(a) seq17:259 (b) seq17:349 (c) seq17:415 (d) seq17:471

Figure 3.2: **(Left to Right)** Frames 259, 349, 415 and 457 of sequence *seq17* in region-based dataset with ground-truth overlay. **a:** A blue hooded truck enters the scene, **b:** The truck gets completely occluded behind the white van, **c:** The truck turns around the roundabout, **d:** The truck comes near the camera before going out of the FoV, it completely occludes every other moving object and experiences perspective distortion as compared to when it first entered the scene.

of long sequences, and especially in the presence of real distortions.

As mentioned in Section 2.3, in contrast to [78], in [45] a human-inloop methodology was proposed to create a ground-truth motion database in both indoor and outdoor natural scenes. The authors designed an interactive publicly available computer vision system to allow a user to efficiently annotate motion. The limitation in this tool was that it was not designed to annotate long sequences with real distortions. The propagation of region label fails when a moving object becomes occluded, or when the motion region boundary does not adjust as the object's boundary expands or retracts due to perspective distortion. These limitations were difficult to handle as it meant manual correction in each frame, but certain features of the tool made it possible to use it while not letting computational time overhead become too large.

In rigid objects, there were a few scenarios which were particularly difficult for annotation, explained in Fig. 3.2. During complete occlusion, region boundary propagation of the blue truck stops, seen between Fig. 3.2a and Fig. 3.2b, as the tool assumes that the object has gone out of the FoV. This phenomenon is overcome by using the replication tool over occluded frames. In the frames around Fig. 3.2c the object takes a turn, which is difficult to annotate as there is a considerable change in shape. The region boundary in each frame while turning motion is replicated and shape corrected for correct annotation. The object's appearance, changes a lot along its motion from Fig. 3.2a to Fig. 3.2d due to perspective distortion. During this transition, the region propagation is done piece-wise in steps of 5 frames at a time.

In the case of non-rigid and articulate objects i.e. people, the propagation is done piece-wise like in the case of perspective distortion. The label is propagated over a small number of frames assuming that the motion remains consistent during those frames. Then the region boundary is corrected and then re-propagated. This is repeated for each non-rigid object till it remains in the field of view. The completion of this annotation process results in 235 annotated motions in 20 long sequences. A log of annotation time taken for each motion in each sequence is also kept and provided with the database, so that if the ground-truth of this dataset is used in the domain of label propagation, then a quantitative analysis is possible.

3.3.2 Region-based short sequences

As mentioned in Section 3.2.2, most of the state-of-the-art algorithms cannot deal with long sequences as they are designed to be tested on Hopkins155 dataset. They are designed to only evaluate short sequences with no missing data. This problem is again solved by creating a region-based short sequences dataset, from the long ones. The length is shortened by cutting the long sequences into multiple short ones, making sure that real distortions are still prevalent in each small sequence.

The ground-truth regions of all the short sequences are taken from their corresponding original region-based long sequence ground-truth frames. The sequence cutting leaves a residue of less representative motions in some short sequences, which are kept unlabeled. For each short sequence, a trajectory-based ground-truth formed from its respective region-based dataset is also given. In total, 150 short sequences are created from 20 long sequences. Each short sequence contains all or a combination of real distortions. The details of all the short sequences can be seen in Table 3.1, where 150 short sequences contain 440 motions, excluding background camera motion, spanning over 6200 frames. A further subset of 34 short sequences with no missing data is also created to process state-of-the-art algorithms, which are unable to handle missing data.

All the dataset resources including sequences in original and reduced resolution, trajectory- and region-based ground-truths, evaluation source codes, results of the benchmarks and the related documentation are publicly available at `http://dixie.udg.edu/udgms/`.

3.4 Motion segmentation benchmark

3.4.1 Benchmarking methods

The choice of an evaluation criteria in such diverse database is critical. The criteria should provide an insight into an algorithm's performance on both, trajectory- and region-based datasets. To achieve this task, we follow the proposal similar to the one presented in [39]. If the classifier produces dense region labels, then the application is similar to the one proposed in FBMS59 [39]. If trajectory labels are to be adjudged, then there is an underlying assumption that the order of classified trajectory labels is the same as the order of trajectories in the given trajectory matrix. Based on this assumption, all the intermediate and final measures are resolved. As most MS methods are evaluated using trajectories [31,33–35,38,43,55,56], it is critical to carefully confirm that the assumption holds for the classified labels.

Considering N ground-truth labels (GT) and M segmentation labels (SL), a label correspondence matrix, $LC_{N\text{x}M}$ is computed. The evaluation measures true positives (TP), false positives (FP) and false negatives (FN) estimate *Sensitivity* $S = TP/(TP+FN)$, and *Precision* $P = TP/(TP+FP)$ as in [39]. With S and P known, we compute the F-measure (F score), using Hungarian method as in [39],

$$F_{i,j} = \frac{2 * S_{i,j} * P_{i,j}}{S_{i,j} + P_{i,j}} \tag{3.1}$$

The F-score is computed for each pair of GT_i, $i = 1...N$ and SL_j, $j = 1...M$. The best assignment of SL with GT is found by maximum F-score, and is stored in the label correspondence matrix, *LC*. If N>M, there remain unassigned GT labels, and if N<M, there are unassigned SLs.

Table 3.3 and 3.4 show all the performance measures of the state-of-the-art benchmarking algorithms for both datasets. An **average region density measure D** [39], which is the average percentage coverage of labels over all ground-truth regions, is used. This is valid only for region-based benchmarks. There are three measures for misclassification; *Overall misclassification, Motion separation and Misclassified motion labels*, whose ideal percentage value is 0. An F-measure, whose ideal value is 100% and a threshold based estimation of number of motions.

- **Overall Misclassification:** The Overall Misclassification (OM) is the total number of misclassified trajectories in the sequence computed based on the best correspondences found in *LC*. It is the standard measure used in Hopkins155 benchmark [38]. OM is misleading in our case because it is highly biased, as a large percentage of trajectories belong to the background. So, if the algorithm undersegments and classifies everything as background, it can have very low OM. Therefore, two different measures, which ignore all background trajectories, are also computed.

- **Motion Separation:** Motion Separation (SM) is the percentage of motion that was wrongly assigned to the background by an algorithm. This measure specifically gives an insight on the MS algorithm's ability to separate motion from background, without getting into the correctness of the assigned motion label.
- **Misclassified Motion Labels:** Misclassified Motion Labels (ML) is the percentage of wrongly assigned motion labels computed based on the best correspondences found in *LC*. This measure gives an insight on the actual performance of motion segmentation.
- **F-measure:** It is the average F-measure, which refers to the harmonic mean of *Sensitivity* and *Precision*. It is the measure proposed in FBMS59 [39] for performance evaluation. It gives an overall view of segmentation of motion by the algorithm.
- **Extracted regions:** The number of extracted motions are reported by the same criteria, F-measure $F \geq 75\%$, as in [39], to maintain continuity.

3.4.2 Experimental setup

The challenges present in these motion segmentation datasets, especially in long sequences, render most state-of-the-art MS algorithms unusable. Nevertheless, OB algorithm [65] is able to attempt these challenging long sequences. Therefore, a detailed evaluation of OB algorithm on long sequences of both datasets is presented in Table 3.3.

The choice of algorithms to benchmark our dataset is quite difficult. There are two main factors effecting the selection of an algorithm; generality in application and availability of the source code. Several algorithms introduced in Section 2.2 are deficient in one of these two terms. For instance, the approaches presented in [37, 64, 68] lack generality of application. The work introduced in [37] is limited to human motion segmentation, the work presented in [68] is applicable only on self-occluding motion, while the approach proposed in [64] is for crowded people scene specific MS, with an additional input requirement of annotated silhouettes masks. Two other approaches [25, 63], whose code is available, have other limitations. The work in [63] has a limitation in

Table 3.3: Benchmark evaluation results. Acronyms are **OB:** Ochs *et al.* binary in [65], **wBG:** with background labels, **woBG:** without background labels, **TBLS:** Trajectory-based long sequences, **RBLS:** Region-based long sequences, **SL:** # classified labels, **OM:** overall misclassification, **SM:** separation of motion from background, **ML:** misclassification of motion labels w/o background.

Ochs-Brox (OB) algorithm							
Name	D(%)	SL	OM(%)	SM(%)	ML(%)	F-measure(%)	$F \geq 75\%$
TBLS wBG	–	154	17.66	34.01	48.90	42.80	72/201
TBLS woBG	–	132	–	–	34.07	46.55	66/181
RBLS wBG	0.69	206	27.04	25.99	43.39	43.79	92/235
RBLS woBG	0.69	179	–	–	30.36	49.68	95/215

application as it deals with only one object per frame, while the work presented in [25] is heavily reliant on optical flow and occlusion detection correctness. Besides, for some other methodologies [61,62,67], which are generally applicable, authors have not shared their source code for testing. Hence, the selection of benchmarking algorithms presented in this work is done based on these constraints.

In order to demonstrate the difficulty of the challenge, a detailed analysis of the methodologies, which are applicable on small sequences datasets, is also presented. The best-performing MS algorithms on other datasets, with source code availability and general applicability, are chosen for evaluation on our database. We evaluate the small sequence datasets, with and without missing data, on the factorization based method (ALC) [43], which can deal with missing data. We also present the evaluation on the enhanced subspace affinity based method [34] (ELSA), and on the current Sparse Subspace clustering method (SSC), with its low rank (LRR) [32, 33], and latent space (LS3C) [44] variants. Moreover, the results of OB algorithm on small sequence datasets is also presented in Table 3.4.

A 64-bit Intel i7 core 3.4 GHz machine with 16GB RAM was used for processing, except OB which was run in a similar server machine with 128GB RAM. All the scripts and results related to the benchmark experimentation are publicly available online with the datasets. The scripts are also designed to be able to incorporate any new algorithm for standardized comparison of results on our datasets.

3.5 Experimental results

3.5.1 Quantitative results

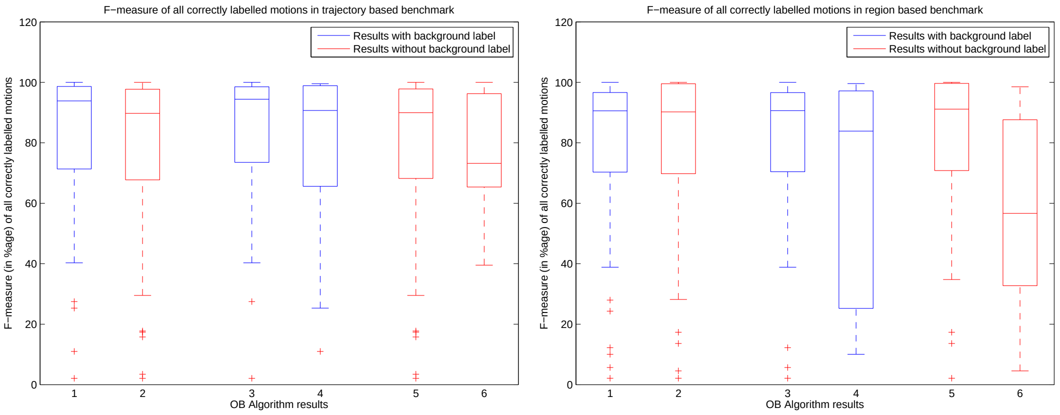
The results in Table 3.3 clearly show that difficulty of the dataset on long sequences is considerably increased. The average F-measure, which in the case of FBMS-59 was 70% on the training set and 65% on the test set [39], is down to 42.8% on our trajectory-based benchmark and 43.8% on the region-based benchmarks.

The SM and ML measures in both benchmarks indicate that the algorithm performs a better separation of motion from background than intra-motion classification. OB fails to recover motion after complete and multiple occlusion. The number of classified labels SL are well below the number of ground-truth labels, which clearly shows under-segmentation. This occurs due to the failure of separation of motion from background and poor segmentation separation of similarly moving separate objects. OB also does not deal with camera motion effectively, as the yaw motion of the camera results in

Table 3.4: Benchmark evaluation results. Acronyms are **OB:** Ochs *et al.* binary in [65], **SL:** # classified labels, **OM:** overall misclassification, **SM:** separation of motion from background, **ML:** misclassification of motion labels w/o background. **F.m.:** F-measure, **PT:** processing time (in seconds).

Trajectory-based short sequences (TBSS)												
Algos	**Short sequences with no missing data**						**Short sequences with missing data**					
Name	D(%)	SL	ML(%)	F.m.(%)	$F \geq 75\%$	PT	D(%)	SL	ML(%)	F.m.(%)	$F \geq 75\%$	PT
LS3C	–	–	26.69	64.40	48/101	89	–	–	41.08	44.10	94/442	2282
SSC	–	–	20.73	**83.05**	68/101	3963	–	–	41.17	38.89	87/442	16181
ALC	–	66	34.18	49.15	34/101	11797	–	571	29.90	55.23	148/442	35711
ELSA	–	**93**	**13.08**	82.20	**79/101**	180	–	507	42.94	48.42	90/442	2444
LRR	–	414	41.32	48.15	26/101	**79**	–	1589	42.18	36.66	83/442	**213**
OB	–	90	17.33	72.04	66/101	–	–	**424**	**16.36**	**69.77**	**284/442**	–

Region-based short sequences (RBSS)												
Algos	**Short sequences with no missing data**						**Short sequences with missing data**					
Name	D(%)	SL	ML(%)	F.m.(%)	$F \geq 75\%$	PT	D(%)	SL	ML(%)	F.m.(%)	$F \geq 75\%$	PT
LS3C	0.75	–	26.84	61.83	41/81	**84**	6.21	–	43.14	43.57	90/475	1054
SSC	0.75	–	20.16	83.98	58/81	2666	6.21	–	44.13	37.66	73/475	16016
ALC	0.75	64	19.10	61.45	39/81	2374	6.21	1044	51.74	47.10	84/475	28275
ELSA	0.75	**89**	**12.29**	**85.37**	**67/81**	141	6.21	**482**	46.00	46.10	87/475	2828
LRR	0.75	286	43.57	44.24	19/81	207	6.21	1345	45.42	33.27	64/475	**252**
OB	0.75	68	19.88	70.69	51/81	–	6.21	420	**18.86**	**65.82**	**272/475**	–



(a) Trajectory-based benchmark

(b) Region-based benchmark

Figure 3.3: Boxplots of F-measures of all the correctly labeled motions in the each benchmark. Results are with or without background labels and with or without relative motion. **Boxplots 1 and 2** contain all sequences, **Boxplots 3 and 5** contain sequences without relative motion, **Boxplots 4 and 6** contain sequences with relative motion.

multiple labels on the non-homogeneous regions of the background, as shown in Fig. 3.4.

The sequences with relative motion, when the camera is inside a moving car, present a greater challenge as the algorithm performs poorly on them. In Fig. 3.3a and Fig. 3.3b, boxplots 4 and 6, it can be seen that min-max F-measure range of only the detected motions in sequences with relative motion is quite high. This is because the OB algorithm uses LDOF as its core information, and in the presence of relative motion this criterion becomes increasingly noisy. The absence of background label increases the confidence of motion detection, resulting in compact box plots and high F-measures. Fig. 3.3a and Fig. 3.3b, boxplots 1 and 2, also depict that when the motion is correctly segmented, it is segmented with a high confidence, which makes OB robust.

The results on small sequences, see Table 3.4, depict similar results as in the case of long sequences benchmark. Only SSC and LS3C have the number of motion clusters as a-priori knowledge. In both the short sequences datasets with no missing data, OB and ELSA perform slightly better than the rest. The misclassification of the algorithms predominantly ranges up to 25% even in results with no missing data, which shows the difficulty in the separation of motion patterns in our datasets.

The overall results on short sequences with missing data are also similar to the ones on long sequences. Only OB and ALC are able to deal with missing data, so both outperform other algorithms. As OB is able to robustly recover motion labels of partially occluded moving objects, it has the minimum misclassification, maximum F-score and the highest number of correctly detected objects in both benchmarks. ALC

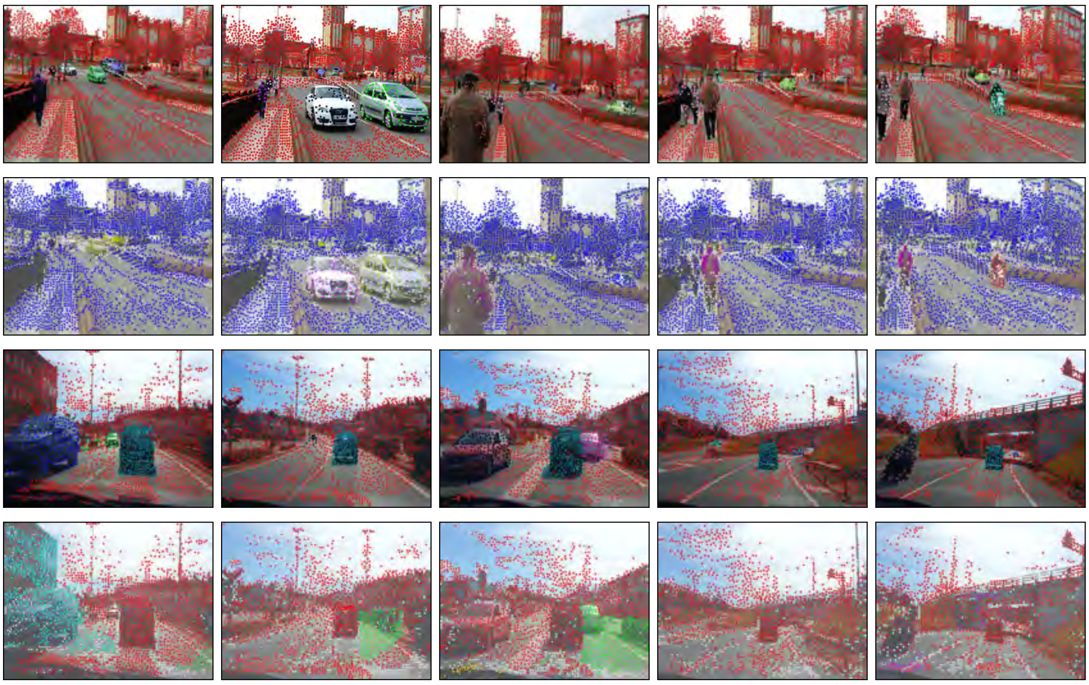


Figure 3.4: **(Left to Right) (Row-1 and Row-2)**: Frames 600, 693, 816, 939 and 1024 of sequence *seq14* of trajectory-based long sequences benchmark. **Row-1:** Images with trajectory label ground-truth overlay, **Row-2:** Images with OB algorithm result overlay. **(Row-3 and Row-4)**: Frames 48, 217, 286, 689 and 744 of sequence *seq11* of trajectory-based long sequences benchmark. **Row-3:** Images with trajectory label ground-truth overlay, **Row-4:** Images with OB algorithm result overlay.

also performs well but its processing time is extremely high. Even the best performing algorithm is able to achieve 65% F-measure, which is an evidence regarding the difficulty level posed by our dataset.

It is pertinent to observe that the result of all the methodologies on our benchmarks is consistent with their results on the previous benchmarks. Therefore, considering the resulting measures on an equal scale, we can safely say that our benchmark of long, as well as of small, sequences poses an increased and diverse challenge for the community.

3.5.2 Qualitative results

The qualitative results show that our benchmarks contain intricate challenges, which were not captured in any other dataset yet. Fig. 3.4 illustrates the OB algorithm results for 5 frames each of *seq14* and *seq11* in trajectory-based long sequences benchmark.

In *seq14* results Fig. 3.4 (Row-2), it can be observed that with minimal or no camera motion, the background labels remain consistent. The relative changes in size during motion, from small to large or vice versa, due to the position of the object with respect to the camera or even due to perspective distortion, does not effect the segmentation result. Similar, spatially separable, parallel motions are given separate labels. It was observed that this behavior of the algorithm depended on the speed of motion. Small objects with considerable motion but far from the camera are not segmented as tracking fails on these objects. Non-rigid motion of people is segmented as long as densely populated trajectories are present. On the other hand, inaccuracies are generated in the segmentation of sparsely distributed trajectories, which are not representative enough of their respective non-rigid motion. Motion labels after complete occlusion are not recovered, as there is no post-occlusion trajectory recovery or matching mechanism present in any algorithm.

In Fig. 3.4 (Row-4), *seq11* results highlight that relative motion is extremely difficult to segment, especially in an optical flow based method. All objects having small relative motion are merged with the background. This can be a desired effect in the foreground detection of large objects. When there is a camera motion transition, from fast to slow or vice versa, then the whole region, with the moving object present in it, is segmented as one object. When the relative motion difference is large, then the object is segmented but along with its spatially attached background region.

Fig. 3.5 illustrates the OB algorithm results for 5 frames each of *seq14* and *seq11* in the region-based long sequences benchmark. In Fig. 3.5 (Row-2), *seq14* results depict that a panning motion of camera gives rise to multiple labels on background. Either a stop-start camera motion, or a static object crossing extending across the frame, generates a new background label, which is undesirable. Each multiple and complete occlusion of an object results in a new label for this object. Objects are well segmented in general, but those covering a small spatial region after complete occlusion, are sometimes merged with the background.

In Fig. 3.5 (Row-4), the results of *seq12* with a tower camera view are illustrated. The sequence is easy as there are few or no occlusions, which makes trajectories or moving object regions consistent and easy to segment. In this case a small rotation of the camera results in unnecessary splitting of the background label. Objects with similar motion, in a line or a curve, even though spatially separable, are not segmented. Stopping motion is not segmented well if the object movement is not sufficiently large. Only the last car in the last frame with sufficient movement after stopping motion is segmented, all the other cars in front of it are merged with the background.

OB performs best in small sequences as expected, because it is able to deal with missing data. It is also able to recover the label of partially occluded objects. The cost

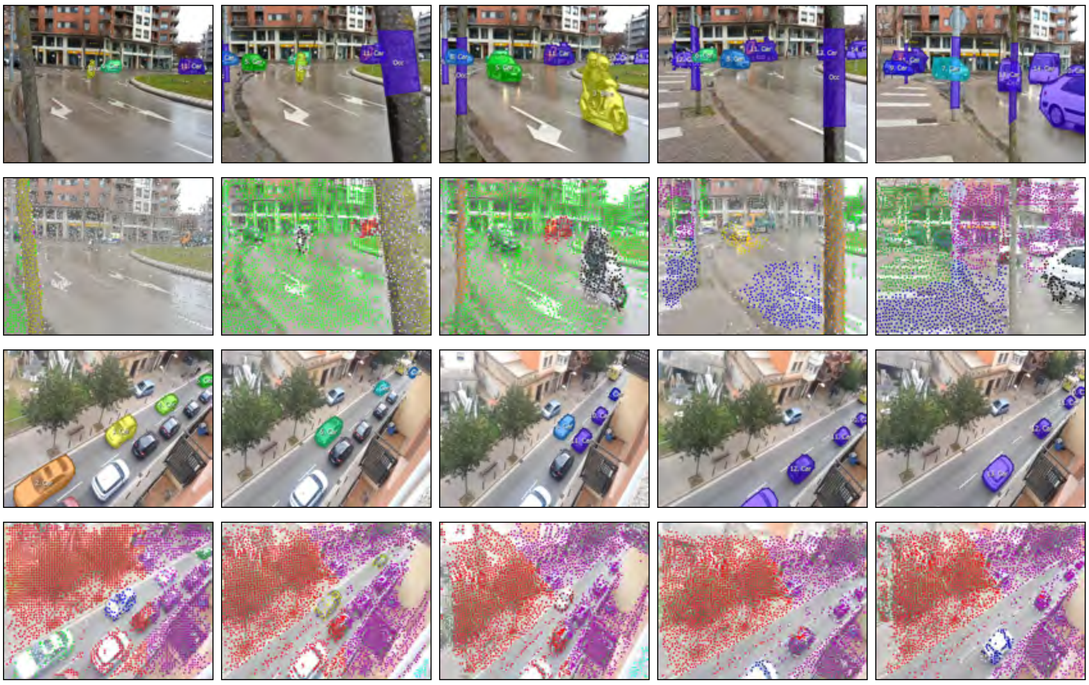


Figure 3.5: **(Left to Right) (Row-1 and Row-2)**: Frames 116, 157, 200, 257 and 351 of sequence *seq14* of region-based long sequences benchmark. **Row-1:** Images with region label ground-truth overlay, **Row-2:** Images with OB algorithm result overlay. **(Row-3 and Row-4)**: Frames 107, 394, 572, 670 and 762 of sequence *seq12* of region-based long sequences benchmark. **Row-3:** Images with region label ground-truth overlay, **Row-4:** Images with OB algorithm result overlay.

function in OB, inherently looks in to shape, color and texture features, because of this even in the presence of partial occlusion, the motion label remains consistent. Other algorithms lack in this capability of handling missing data and real distortions, and are therefore unable to provide a good segmentation result.

3.6 Conclusion

In a nutshell, the presented datasets can be treated as a consolidation of all the dataset varieties available in the current state-of-the-art. The ground-truth variations, sequence length variabilities and complex motions spanning over long frame lengths pose an enhanced challenge for the MS community. Nonetheless, to exactly determine the place of the presented datasets among the current state-of-the-art contributions, its similarities, differences and limitations are listed in Table 3.5.

Based on the given attributes, the presented dataset can be placed as a benchmark, which specializes in real complex rigid motions, with outdoor street background, undergoing multiple distortions, along with an added variation of non-rigid motion of people. The moving objects contain variations in depth, illumination, stopping motion, occlusion, speed, lightning conditions and shadows. Despite the presented features of the dataset, there are certain limitations in it as well. The overall number of non-rigid motions are less, more specifically some type of classes are not addressed. The non-addressed classes include animals motion, birds motion, airborne things (planes, gliders, parachutes, etc.) and underwater motions. These classes are application specific, therefore, they can be added at a later stage to consolidate the dataset.

In spite of the absence of some non-rigid classes, the most challenging task in the making of such datasets is motion annotation. In ur presented datasets, the motion annotation was done for motion-trajectories as well as motion-regions. A state-of-the-art of motion annotation was described in Chapter 2, based on which we put forth two proposals in Chapter 4.

Table 3.5: The comparison of dataset varieties. Their similarities, differences and limitations. Acronyms are: **TGT:** Trajectory-based ground-truth, **RGT:** Region-based ground-truth, in Motion types **R:** Rigid, **NR:**, in Occlusions **P:** Partial, **C:** Complete, **Seq.:** Sequences, **S:** Short, **L:** Long, **Var.:** Variation, **HDV GT:** High-Definition Video Ground-Truth, **RM:** Rigid Motions, **NRM:** Non-rigid Motions, **PL:** Planes, **TR:** Trains, **AN:** Animals, **BD:** Birds, **BG:** Background

Variety of Datasets																
Datasets	**Dataset Properties**															
	Similarities				**Differences**								**Limitations**			
	TGT	RGT	Motions (R & NR)	Occlusions (P)	TGT & RGT	Seq. Type (S & L)	Depth Var.	Perspective Var.	Occlusions (C)	Multiple Distortions	No Synthetic Seq.	Stopping Motion	No HDV GT	No RM (PL/TR)	No NRM (AN/BD)	No Indoor BG
Hopkins155 [38]	✓	X	R	X	X	S	X	X	X	X	X	X	✓	✓	✓	X
FBMS59 [39]	X	✓	✓	✓	X	✓	X	X	X	✓	✓	X	✓	✓	X	✓
SegTrack [52]	X	✓	NR	✓	X	✓	X	X	X	✓	✓	X	✓	✓	X	✓
MOViCS [51]	X	✓	NR	✓	X	L	✓	X	X	✓	✓	X	✓	✓	X	✓
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

> Labels are for filing. Labels are for clothing. Labels are not for people.
>
> Martina Novratilova

Chapter 4

Tools for motion annotation in video sequences

Motion analysis is a pre-requisite in video analysis with its applications in many domains of computer vision. Objective analysis of moving objects can be carried out when motion is accurately detected and segmented as a prior. In the state-of-the-art of computer vision, precise and robust algorithms, which can work in the presence of occluders and other distortions, while the acquisition of video is done from a moving camera, are still elusive.

4.1 Motion annotation problem

The limitations prevailing in annotated moving objects' datasets are restricting the development of effective motion analysis tools. The diversity and complexity of a real life motion captured in a collection of video sequences, determines how representative the dataset is of the actual problem. If the annotated datasets encapsulate limited motion diversity, then the algorithms tested on them will also have limited applicability. On the other hand, if more complex motions are captured in a sequence for dataset formation, the dataset will become more representative. The motion complexity makes the task of correctly generating ground-truth motion label for each moving object in all the frames of a video sequence increasingly cumbersome. This task of generating the ground-truth label for each motion on all the frames of all the video sequences in a dataset is known as the motion annotation problem.

If seen with respect to the trajectory- and region-based datasets presented in Chapter 3, the motion annotation can be resolved into two categories:

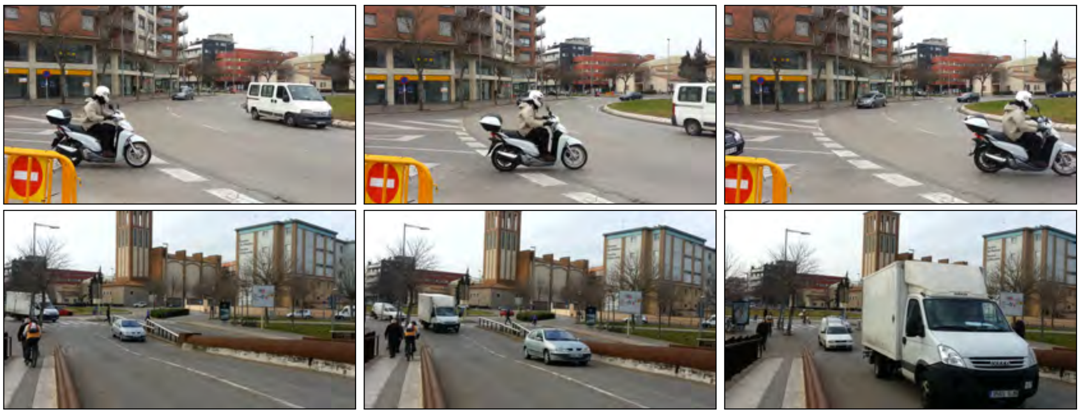


Figure 4.1: First, middle and last frame of a moving object while entering and leaving the field of view in a video shot. **Top:** The bike enters and leaves the frame without occlusion and distortion. **Bottom:** The white truck enters and leaves the frame while undergoing partial occlusion, change in heading direction and illumination, a significant alteration in relative size and experiences perspective distortion.

- **Motion-trajectory annotation**: It is the task of generating unique motion labels on pre-computed trajectories capturing all the motions in a video sequence. This task was performed to create the trajectory-based long sequences dataset as explained in Section 3.2.1.
- **Motion-region annotation**: It is the task of generating unique labels on all the moving object regions in a video sequence. This task was performed to create the region-based long sequences dataset as explained in Section 3.3.1.

In the motion annotation problem, the problematic element is the expert-user annotation time, which increases many-fold as the captured motion becomes excessively complex. An illustrative example is presented in Fig.4.1, which shows the first, middle and last frames of two moving objects in a video shot, while they enter and leave the field of view. The biker in the top row remains unoccluded, relative change in size across all frames is minimal, the illumination remains generally homogeneous and no perspective distortion effect can be seen. On the other hand, the white truck, present in the bottom row, enters the field of view with a small size due to being considerably deep in the scene with reference to the camera, experiences partial occlusion during the course of its motion, and exits the frame with an enlarged size, change in heading direction, variation in illumination and with perspective distortion. The expert-user annotation time for generating ground-truth on these two motion samples is radically different. While the annotation labels on the white car may be generated with state-of-the-art label propagation algorithms, there is no modern, time efficient methodology or platform, to annotate the white truck or such motions. Though, this example elaborates the complexity entailed in the motion-region annotation problem, the same holds for the motion-trajectory annotation. As more complex motions are captured in a video sequence, the associated tracked trajectories become more complex and noisy. Annotating such noisy trajectories in an automatic or a semi-automatic way becomes difficult.

These impending issues of motion-trajectory and motion-region annotations can be resolved through task specific annotation tools. The requirement of such tools are that they should be user friendly and time efficient. Two such proposals, one for each modality are proposed in the following sections.

4.2 Tools for motion trajectory annotation

The most known trajectory-based publicly available datasets is Hopkins155 [38]. It still remains the most extensively utilized MS dataset. This is so because of the inherent sparse trajectories present in it and the availability of a standard evaluation metric, which made it easy for algorithms to be compared at an equal scale. Since, Hopkins155 further trajectory-based datasets were not proposed because of the inherent difficulty in annotating complex trajectories.

4.2.1 Trajectory annotation tool

In a trajectory-based dataset, given a video sequence, the moving objects are tracked through sparse trajectories to build a trajectory matrix $W_{2f \times p}$, where f is the number of video frames and p is the number of tracked feature points (trajectories). The sparsity of the tracked feature points on the moving objects varies with respect to the desired density of coverage. Independent of the density of coverage, a trajectory-based dataset with long real sequences can only be created in the presence of a trajectory annotation platform. 'Trajectory annotation tool (TAT) is one such platform, where complex trajectories can be annotated.

TAT is developed using MATLAB libraries for Graphical User Interface (GUI), with all the resources publicly available on-line at `http://udgms.udg.edu/TAT/`. The scheme of the TAT annotation tool is presented in Fig. 4.2, with each process pipeline inside. TAT has four major processes, *Label initialization*, *Data update engine*, *Display engine* and *Save workspace.*

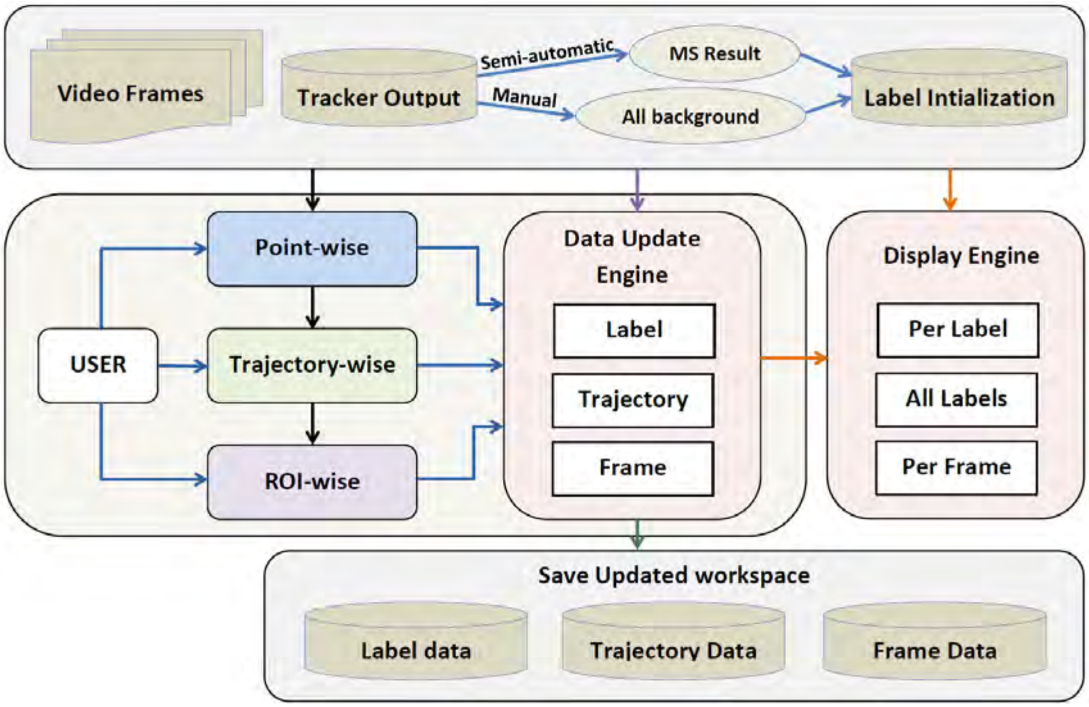


Figure 4.2: TAT flow-diagram, input block at the top with label initialization modes, data update and display engines in the middle with save option at the bottom. User can label through three modalities: Point-wise, Trajectory-wise and ROI-wise. The display engine can exhibit annotations in three forms: Label display per label, all labels or per frame.

4.2.1.1 Label initialization

The inputs of the system are loaded in the Label initialization module.The inputs are the frames of a video shot, which is to be labeled. These frames are fed in a point tracker algorithm, LDOF [46], which gives a Trajectory matrix, $W_{2f\times p}$, as output. All the moving objects' and the background motion are encapsulated this W matrix. In W, each column represents the number of points in each frame, where as each pair of rows represent a complete trajectory. An input *Label vector* $L_{p\times 1}$ is also used, which changes according to the selected initialization mode. The TAT tool offers two modes of annotation in the workspace, where each starts with its own initialization. The two modes are,

- **Semi-automatic**: In the semi-automatic mode, the OB [39] MS algorithm is used as label initialization, and then the annotation is refined to form a final ground-truth. It should be noted that TAT provides the flexibility to use any MS algorithm for initialization.

- **Manual**: In *manual* mode, the label initialization is done by annotating all trajectories as background by default.

When using the semi-automatic mode, the output of the OB MS algorithm is used as a label vector, while in manual mode, all the trajectories are initialized with the background label, by default.

4.2.1.2 Data update engine

After mode selection and label initialization, the initialized label data is passed to the Data update engine. It takes the input from the user to rectify the labels's of the trajectories, which are incorrectly labeled. The user is provided with three point/trajectory selection methodology options to update the labels. These selection methodologies are,

- **Point-wise:** This option facilitates point by point selection of trajectory data points of the active frame on the display window. In this mode, the whole trajectory of the select point can be visualized, if the 'Current Frame' option in the 'Display Setting' is selected.

- **Trajectory-wise:** In this option, upon selection of one point on a trajectory a complete trajectory gets selected. If an assignment of a new label is made, then the whole trajectory gets labeled. If multiple points on the active frame are selected, then each trajectory corresponding to each point gets selected and labeled.

- **Region of interest (ROI)-wise:** In this option, a rectangular bounding box shaped ROI is available to the user to select a perform labeling on a cluster of trajectories in one go. On the single active frame, all the trajectories of the selected points inside the user drawn ROI get labeled.

In the Data update engine, these input variables are converted into three indexes: label, trajectory and frame. Each index is dependent on the annotation update.

1. **Label index:** The label index uses the label vector $L_{p\times 1}$ to keep track of the ID tagged with each motion independent of the initialization mode. When a trajectory on a new moving object is selected and annotated, a new label is added in the label index. As more trajectories on a motion are annotated, they form a trajectory cluster of the object across all the frames. If at any point during annotation, an already existing label ID is given to a new trajectory cluster then

the two clusters are merged. This provision is specially useful if, mistakenly, two trajectory clusters with unique IDs are formed of the same moving object then both can be merged in one go.

2. **Trajectory index:** The trajectory index is formed by using the trajectory matrix with the label index. It keeps a log of association of each trajectory with its annotated label index ID. This is specifically useful in the display engine as all the trajectories belonging to a unique label index ID are displayed in a unique color. The trajectory index is essentially used to form the annotation result, *Updated Label vector* $Lu_{p\times 1}$.

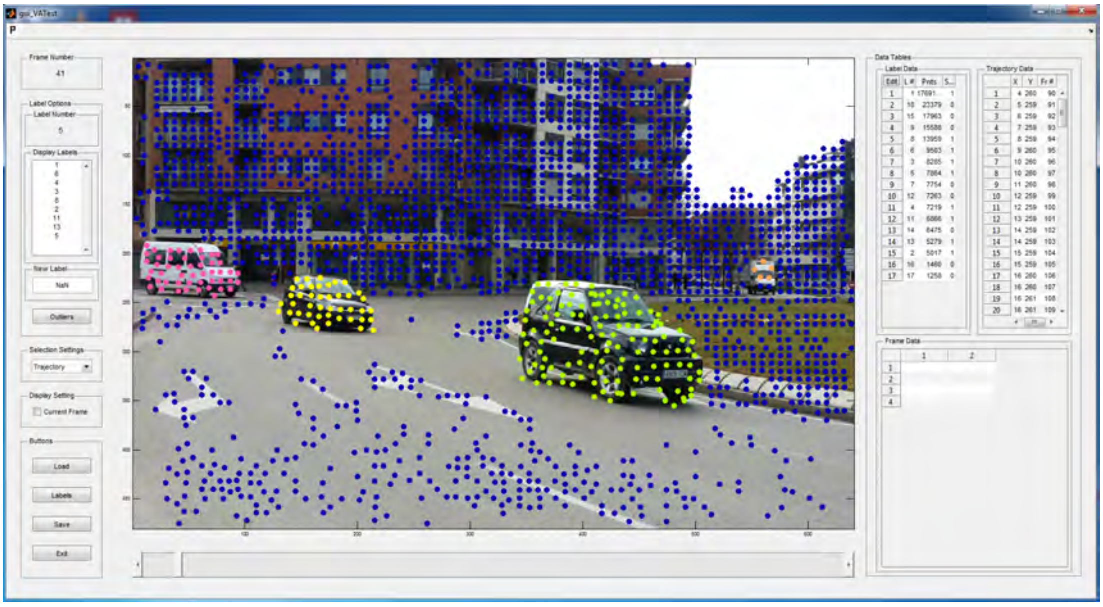
3. **Frame index:** The frame index is formed by the input frames of the video sequence along with the trajectory index and the label index. This index is the back bone of the display engine, which facilitates swift annotation in TAT while providing a visual check on the correctness of each annotated motion at the same time. A log of each trajectory with reference to its updated label present in each frame is kept in this index. The index contains multiple trajectory clusters with associated labels per frame. While scrolling through the frames this index shows each label ID with reference to the selections made in the display engine modalities.

In TAT, these indexes provide the structure for data update and display engines to be efficiently used. The display window of the GUI is shown in Fig. 4.3, where a frame with trajectory overlay, after complete annotation on a few moving objects, can be seen.

4.2.1.3 Display engine

The display engine exhibits the frame in the display window, whose respective trajectories are being annotated. A scroll pane below the display window facilitates scrolling through the frames of the video shot. The frame number of the active frame is displayed on the top left corner of the GUI window. Right below it, the current annotation label of the selected trajectory or point is shown.

Below the label number display, there are two user input boxes. In the first box, the label numbers of the trajectories that the user desires to see on the active frame can be listed. This results in the display of the point on the trajectories, which lie on the active frame. The second box takes the label number input from the user to be assigned to the selected trajectories. If a new label number is given by the user, which



Frame Number: 41

Label Number: 5

Display Labels: 1, 6, 4, 3, 8, 2, 11, 13, 5

New Label: NaN

Selection Settings: Trajectory

Display Setting: ☐ Current Frame

Label Data

Edit	L #	Pnts	S...
1	1	17691	1
2	10	23379	0
3	15	17963	0
4	9	15588	0
5	8	13959	1
6	6	9583	1
7	3	8285	1
8	5	7864	1
9	7	7754	0
10	12	7263	0
11	4	7219	1
12	11	6886	1
13	14	6475	0
14	13	5279	1
15	2	5017	1
16	16	1460	0
17	17	1258	0

Trajectory Data

	X	Y	Fr #
1	4	260	90
2	5	259	91
3	6	259	92
4	7	259	93
5	8	259	94
6	9	260	95
7	10	260	96
8	10	260	97
9	11	260	98
10	12	259	99
11	12	259	100
12	13	259	101
13	14	259	102
14	14	259	103
15	15	259	104
16	15	259	105
17	16	260	106
18	16	260	107
19	16	261	108
20	16	261	109

Frame Data

	1	2
1		
2		
3		
4		

Figure 4.3: The TAT GUI, with display window in the middle, labeling modes and selections on the left and selected indexes on the right.

does not exist in the label index already, then this new label is tagged with the selected trajectories. After each label assignment, all the respective indexes in the data update engine are refreshed. The display engine facilities three modalities,

- **Per label:** This display option takes the label number input from the user. This modality is the same, which was explained in the previous paragraph. The trajectory points of each label number the user lists are shown as overlay on the active frame.
- **All labels:** This option is available in the buttons on the lower left pane. It is activated if the 'Labels' button is pressed. Upon the button press, a display routine is initialized, which displays all the labels per frame. While scrolling through the frames, all the trajectories and respective label overlay can be visualized. Here, the background label is always displayed in blue and the rest of the labels are assigned distinct colors. In one click, the complete annotation progress can be seen. Due to color distinction, any wrongly labeled trajectory with respective to the moving object region boundary can be identified and corrected.
- **Per frame:** This option facilitates the display of trajectories on a single frame. In this modality, the complete trajectory of a selected point of a trajectory on the active frame is shown. The complete trajectory display shows the moving

object motion profile with reference to the FoV. This option can be activated by selecting the 'Current Frame' selection box. The display of a complete trajectory per frame facilitates the visualization of temporal evolution of selected motion, which helps in the identification of noisy trajectories. Noisy trajectories are the wrongly tracked trajectories, which contain motion from two or more moving objects in a single trajectory. Upon identification of a noisy trajectory, it can be made a part of the background or it can be given the label of the motion it is most representative of.

The display engine utilizes the label, trajectory and the frame indexes to exhibit the desired information on the display window. Each display modality facilitates an easy and speedy labeling of the trajectories.

4.2.1.4 Save workspace

The save workspace block is the last process in the TAT's functional pipeline. It saves the user defined ground-truth annotation in two mat files;

- *'seqXX_truth.mat':* Here 'XX' refers to the sequence number. This file contains a structure with four fields containing the information about the video sequence, one field containing the input trajectory matrix and three fields containing information about user annotations.
  - **frames:** Number of frames in the video sequence
  - **width:** Number of columns in each frame
  - **height** Number of rows in each frame
  - **trajectories:** Number of point trajectories resulted by applying LDOF based tracking on the video sequence
  - **W_TrajectoryMatrix:** The input Trajectory matrix, $W_{2f \times p}$
  - **GT_GroundTruth:** The updated ground-truth label vector $Lu_{p\times 1}$
  - **annotationtime:** The total time taken by the user to annotate all the trajectories in the W matrix in seconds.
  - **motionlabels:** The total of number of annotated motions in the sequence including the background label.
- *'seqXX_trajectory_stats.mat':* This file contains a matrix 'statMat' of trajectory-wise statistics per motion label. The file possesses the number of trajectories,

the total number of points in the trajectories and the total annotation time, each statistic per motion label. This file gives a motion-wise breakdown of the annotation task. It gives a localized idea of how easy or complex each motion was.

All the files are saved at a predefined path set by the user in the settings file. The information defined in the settings file is a pre-requisite to the usage of TAT.

4.2.2 Trajectory annotation tool interface

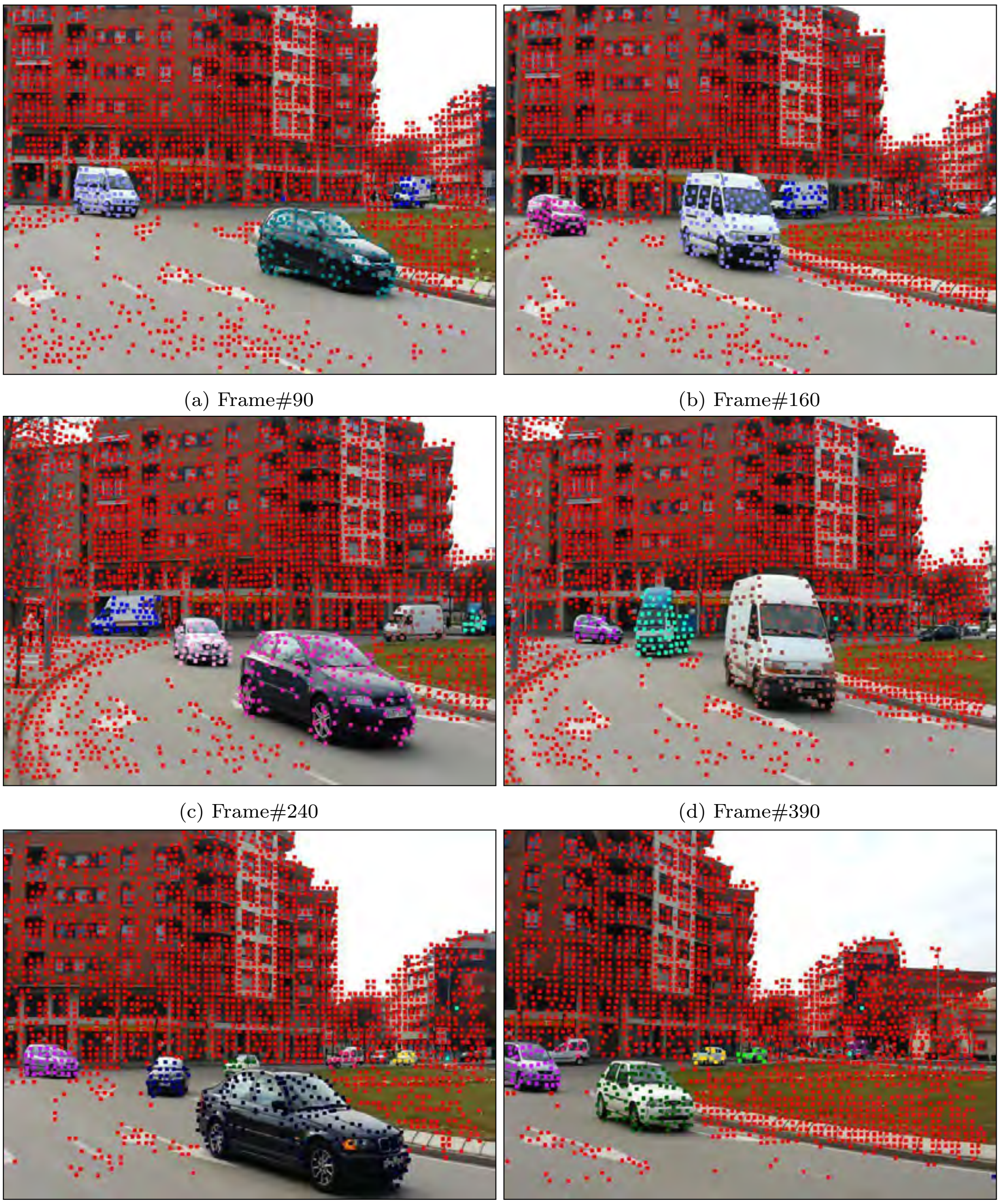
The interface is kept simple with each annotation and display modality directly available for the user. In this way, the tool shares enhanced control with the user so as to maintain flexibility.

In Fig. 4.4, the application of the semi-automatic mode of TAT in a traffic sequence, using OB [39] for label vector initialization, is presented. Depending on the performance of the OB segmentation, the initialized $L_{p\times 1}$ will have some motions, partially or completely, correctly labeled. Although, by using the semi-automatic mode, the overall user annotation time is reduced, the time needed to refine the remaining trajectories is dependent on OB's failures. To reduce the refinement time needed to correct OB failures, the trajectory-wise and ROI wise options are preferable as they select complete clusters in one go. The choice of a trajectory selection option is dependent upon the type (rigid or non-rigid), size and shape of the moving object, its distance from the camera and its position in the field of view. As a thumb rule, the ROI selection option should be preferred in the homogeneous regions of the moving objects. The use of the ROI-wise modality might induce an error in non-rigid, small or irregularly shaped objects, as, neighboring trajectories that belong to either background or other moving objects, can mistakenly get selected. Therefore, for all such objects, trajectory-wise option should be used.

In Fig. 4.4, it can be seen that there are multiple moving object present per frame. Each motion has its own cluster of trajectories spanning over a collection of continuous frames. With reference to Fig. 4.4b and 4.4e, when the moving objects are about to be occluded or are about to enter the FoV, the selection of trajectories on the border of each motion-region becomes critical.

4.2.3 Experimental results

We evaluate the performance of the TAT tool, when creating the ground-truth of the trajectory-based MS dataset explained in Section 3.2.1. This dataset contains 19 long



(a) Frame#90

(b) Frame#160

(c) Frame#240

(d) Frame#390

(e) Frame#510

(f) Frame#600

Figure 4.4: Sample frames of a video sequence in the Traffic group after complete annotation. a) The white vans on either side of the image enter the FoV. b) The size of the white van increases as it comes near the camera. c) Another set of motions enter the FoV with there respective trajectory clusters. d) A blue truck re-appears behind the white van after complete occlusion. e) A white car is being partially occluded by the black car. f) Multiple occlusions about to happen.

video sequences of natural scenes with multiple motions of different types spanning over hundreds of frames having partial and complete occlusions. As compared to Hopkins155, these new challenges presented in this dataset provide a new benchmark for the community.

To quantify the performance of the tool, the total user annotation time, UT is used. This is the cumulative time required to annotate each motion in a sequence. The time taken to completely label each motion is dependent upon the motion type, camera motion, moving object size and frame length of each trajectory. All these traits were extensively tested and the obtained evaluation results are shown in Table 4.11. In order to effectively analyze TAT, the sequences were grouped into four distinct motion types; *Traffic* (MT1), *People-Traffic* (MT2), *Relative Camera Motion-Traffic* (MT3) and *People-Traffic-Camera Jitter Motion* (MT4). The given names exhibit the motion types of each group. We used the F-score [39] to quantify the correctly classified motions by OB initialization. This score takes *Sensitivity* and *Precision* into account. Its scale is in percentage with a maximum of 100%, which would mean that all motion trajectories in a sequence were correctly segmented. To acquire a deeper insight into the annotation time, besides UT, two more time measures were used: UT_s, average time per sequence and UT_m, average time per motion. All the subscripts in our letter denote averages, represented as *s:* per sequence; *t:* per trajectory; and *m:* per motion.

Observing Table 4.11, one can see that it took less UT_s and UT_m to annotate the *MT2* group of sequences. Even though the average frame length of the trajectories in *MT2* were long, this group took less time as it had less motions per sequence, less trajectories per motion and got more than 50% of motions annotated correctly with the OB initialization. UT_s and UT_m were high for the *MT4* group as the motions per sequence were doubled and OB initialization failed on more than 60% motions. Here, less number of trajectories per motion also indicated that the size of objects in the *MT4* group was small, which made the time efficient ROI option unusable. The groups *MT1* and *MT3* took almost the same amount of time on average but the reasons they did so were quite different. In *MT1*, with 62% of motions correctly annotated by the OB initialization, UT should have decreased, but due to more trajectories per motion and more motions per sequence, the overall annotation time increased. In contrast, *MT3* had less motions per sequence so UT should have decreased, however, due to small trajectory frame length and bad OB initialization the overall annotation time increased for these video sequences. The overall average UT was 7.2 minutes in a total of 19 sequences with over 800 frames and 10 motions, per sequence on average.

This evaluation gives an insight about the usage of the annotation tool. It is ap-

Table 4.1: TAT evaluation results. Acronyms are **MT:** Motion Types, **S:** Number of sequences, **M:** Number of Motions, **F:** Number of Frames, **T:** Number of Trajectories, **OB:** OB F-score, **UT:** Total user time, All subscripts are averages; **s:** per sequence, **t:** per trajectory, **m:** per motion.

Trajectory annotation tool evluation results

MT	S	M	M_s	F_t	T_m	OB	UT	UT_s	UT_m
MT1	5	55	11.0	44.5	179.2	62.3	435	87.0	7.9
MT2	7	63	9.0	100.9	126.6	50.8	290	41.4	4.6
MT3	4	34	8.5	42.6	218.9	13.0	275	68.8	8.1
MT4	3	49	16.3	97.6	82.7	38.3	445	148.3	9.1

parent that the semi-automatic modality speeds up the annotation process. The speed up factor depends on $L_{p\times 1}$ initialization, a better initialized label vector results in less annotation time. The ROI option is useful if large non-rigid objects with less occlusion are present in the sequence, as they have motion regions with higher area and density coverage. In small, non-rigid or region borders of objects, it is better to use the trajectory-wise selection option. Though this option is not as fast as ROI in terms of user annotation time, the precision it brings is essential for accurate labeling of these difficult motions.

4.2.4 Conclusion

The creation of ground-truth in trajectory-based MS datasets is a challenging task, especially in the presence of long real-life sequences with multiple motion types and large frame length. Here TAT was presented, which is a semi-automatic trajectory annotation tool for complex videos. It enables the community to create the ground-truth of a motion segmentation dataset on a standardized publicly available platform. We demonstrated that the modalities kept in our tool are flexible, hence the use of any tracker output, and an initialization from any state-of-the-art MS algorithm, is supported. We also provided an evaluation of our tool when it was used to create the annotations on our trajectory-wise long sequences dataset. The evaluation results showed that the platform can produce rapid annotations on long videos with minimal time requirements, which can benefit the MS research community.

4.3 Tools for motion region annotation

The existing methodologies in label propagation address the problem in a limited range of applications. Though, they perform well, they lack utility in real life long videos in outdoor scenes, where multiple occlusions, stopping motion, perspective distortion, multiple appearance-disappearance and noise of camera motion, are present. A reason for these limitations is the absence of a video dataset where these optical phenomenons could be tested. Our proposal in Section 3.3.1 contains these real noises, which makes their quantitative testing possible.

The limitation in label propagation can be looked into as a set of multiple sub-problems based on the complexity and variation in the object motion. The variants include a considerable change in size or illumination, partial or complete occlusion, static or moving occluder, multiple-appearance-and-disappearance in the field of view (FoV), perspective distortion, etc. Each variant, if tackled separately, with a unique approach, can yield improved results. With our work, we aim to tackle these prevalent shortcomings in the label propagation methodology. The results in the state-of-the-art demonstrate that the use of the semi-automatic, as well as the automatic, modality in annotation drastically reduces the expert-user time while preserving the quality of the annotations. We propose a semi-automatic approach by taking annotated labels on two key frames (first and last). We utilize LDOF to promulgate labels across occluders, so that moving object labels are retained even after occlusion. A further refinement of propagated label mask scale is performed by using a non-rigid point set registration method. In this way, we not only improve labels on occluded objects but also in objects undergoing perspective distortion. Furthermore, we provided a consolidated evaluation to establish the usage of our scheme in real life scenes. Our methodology is generally applicable on objects undergoing partial occlusion by static occluders, although it may also be applied on objects undergoing occlusion by other moving objects.

4.3.1 Region annotation proposal

Motion-Region annotation means tagging all the motion-regions with a unique label per motion in a sequence of frames. More formally, given a sequence of N frames $f = \{f_1, f_2, ..., f_N\}$, the objective is to segment all the moving objects M with labels $m = \{m^1, m^2, ..., m^M\}$.

As the goal of annotation is to generate the ground-truth for a given video, it is imperative to take the accuracy of the annotation into consideration. One way of maintaining accuracy is to generate annotation of one motion m^x at a time, with respect to their depth ordering in the scene. The object near the camera first (the one with least depth) and the object furthest from the camera last (the one with most depth). The depth order can be kept track of by the expert user. Hence, the objective is to find the annotation labels m^x where $x = \{1, 2, ..., M\}$, sorted by depth ordering, 1 being least deep and M being the deepest.

The tool presented in [45] facilitates the annotation of moving objects in a sequence of frames. An expert user defines the object outline contour in a key frame. The region inside the object contour is given a label and then the labeled contour is propagated both ways, forward and backward. In the presence of occlusion, perspective distortion and change in object's depth, the propagation fails. If the propagation fails due to illumination variance, background homogeneity with moving object, etc., the labeled region contour can be corrected in frames with bad annotations. The manual correction by the user is then linearly interpolated across all frames the label was propagated on. In the absence of real noise, the platform utilities time efficiently and exhibits good results. On the contrary, it fails in real sequences, especially outdoors, where occlusion, change in depth and perspective distortions are somewhat dominant.

From another perspective, consider the sequence of frames shown in Fig. 4.5 as an example. The moving object enters and exits the FoV in frame f_1 and f_{113}, respectively. The motion annotation of this object, m^1, in these 113 frames can be divided into a set of three sub-problems. One from f_1 till f_{75}, m_1^1, when the object is fully visible without occlusion. Second from f_{76} till f_{100}, m_2^1, when the object is occluded by multiple static occluders. And finally, m_3^1, when the object is again fully visible from f_{101} till f_{113}, until it goes out of the FoV. Then, the overall motion-region annotation of the object, m^x is given by,

$$m^x = \bigcup_{i=1}^{S^x} m_i^x \tag{4.1}$$

In the given example, x, is the object label by depth ordering, and S^x, is the number of sub-problems x^{th} motion-region annotation task was divided into. So, with x being 1 and S^x being 3, the labeled motion-region output of the framework for one object in the given example is given as,

$$m^1 = \bigcup_{i=1}^{3} m_i^1 \tag{4.2}$$

The annotation task of each motion x, to be labeled in the sequence of frames, leads

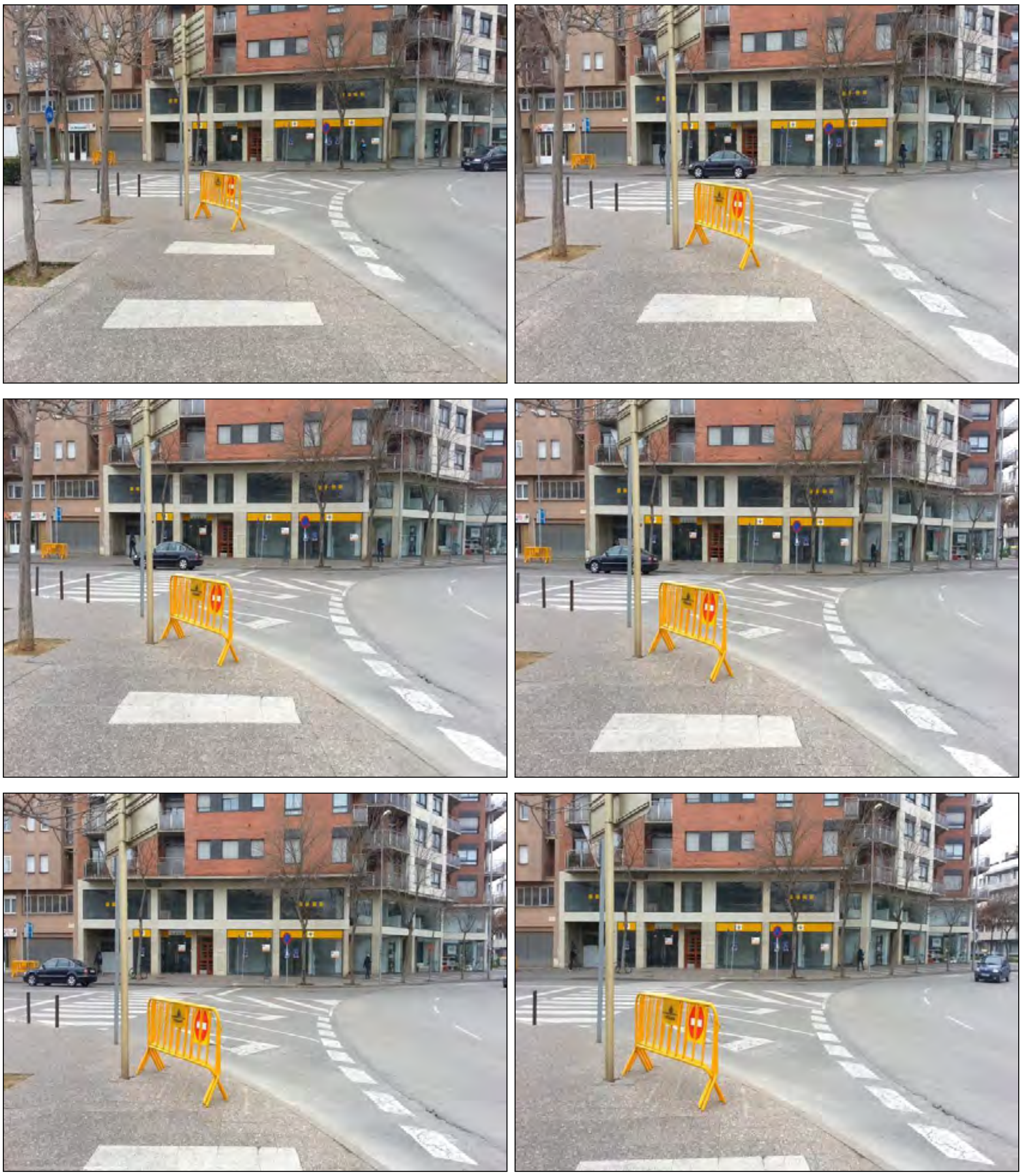


Figure 4.5: Six frames of a moving object, black car, entering and leaving the field of view in a video shot. **Top:** The black car enters the field of view in f_1 (Left) and moves till f_{75} (Right), without occlusion 'm_1^1'. **Middle:** Here, the car undergoes partial occlusion by multiple static occluders from f_{76} till f_{100}, 'm_2^1'. Two frames in this subproblem, where the object was undergoing occlusion are shown, f_{82} (Left) and f_{90} (Right). In f_{82}, the object has started undergoing occlusion behind the two static occluders. In f_{90}, the object has almost gone across the occluders. **Bottom:** The car moves from f_{101} (Left) without occlusion till f_{113} (Right) when it completely goes out of the FoV,'m_3^1'.

to its corresponding sub-problem set S^x.

A modular approach to solve this annotation problem can yield better results in terms of pixel accuracy and time efficiency. This approach of creating sub-problem tasks facilitates the expert-user to objectively divide the annotation problem based on the behavior each moving object exhibits, and also, inherently reduces user annotation time. This sub-categorization based on label propagation complexity can further reduce the manual annotation time, if the label propagation in the problematic subsets, (the ones which require most user corrections due to real distortions), can be automatized.

As mentioned earlier, while [45] works well in unoccluded, low depth change and no perspective distortion motions, it fails otherwise. As a smart hybrid approach, the framework in [45] was used for the subproblems where these distortions were not present. To annotate the subproblems with distortions, we propose a semi-automatic annotation methodology to better utilize the expert-user time. In this section, our annotation algorithm is presented.

Given a set of K frames $f = \{f_1, f_2, ..., f_K\}$, with $K \subseteq N$, in which a single x^{th} moving object appears and then disappears from the FoV. The objective is to find the motion annotation label m^x. It is also given that the annotation task can be further divided into S^x sub problems, where each subproblem can have either of the two types;

- **Type-1 (motion under normal conditions):** Here, the object moves without occlusion or perspective distortion. The annotation under such moving conditions are computed through the work presented in [45].

- **Type-2 (motion under distorted conditions)**: Here, the object undergoes occlusion and/or perspective distortion. The annotation under these conditions are resolved through our motion annotation algorithm.

A pictorial depiction of the same is given in Fig. 4.6. On the top, the figure shows a sequence of N frames, with two moving objects, so the objective is to estimate moving object labels $m^x = \{m^1, m^2\}$. Considering that object 1 is near the camera, it spans over K frames and the annotation task is divided into three subproblems $S^1 = 3$, then $m^1 = \{m^1_1, m^1_2, m^1_3\}$. Here, m^1_1 and m^1_3 are the subproblems of type-1, where the object does not undergo any occlusion or perspective distortion. This annotation problem is estimated by the framework in [45]. On the other hand, m^1_3 is the annotation subproblem of type-2, where the object experiences these distortions. If the movement under distortion spans over L frames, then the objective of the proposed algorithm is to find m^1_3, given the expert-annotated object boundaries in the first and the last frame of the set L. A detailed account of the framework is given below.

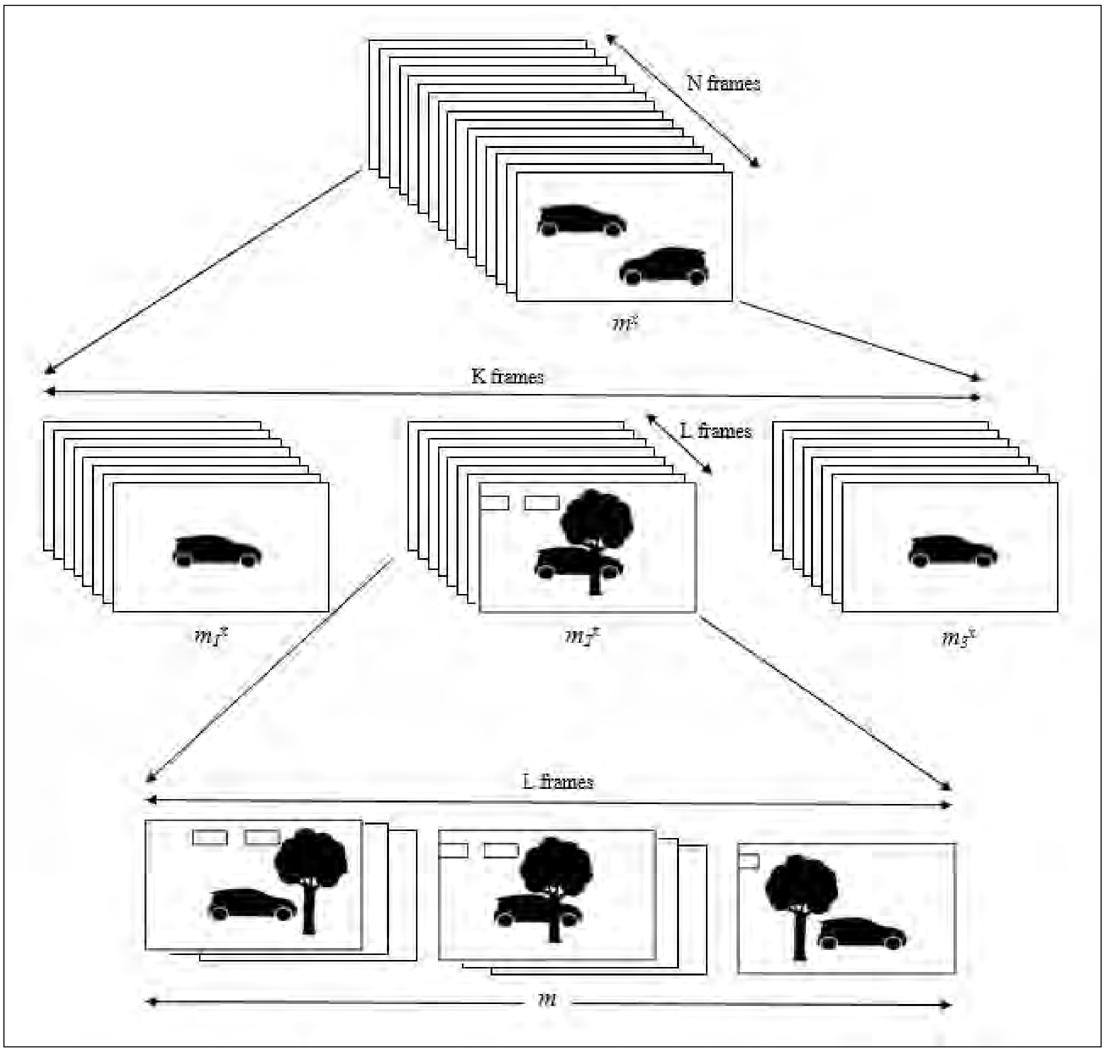


Figure 4.6: Annotation flow of the motion-region annotation algorithm. The motion annotation sub-problem of type-2 spanning over L frames is processed using the proposed algorithm.

In our work, any m_i^x is the output moving object label set computed for all the frames, in sub-problem i of type-2, while annotating moving object x. For ease of notation, any such m_i^x in the remaining text is denoted as m. In our framework, a three pronged motion-region label propagation approach was taken to attain maximal accuracy with minimal expert-user intervention. The steps include *Occluder mask tracking* (m^{occ}), *Object mask propagation* (m^{ini}) and *Object mask scale adjustment* (m). A block diagram of the algorithm is shown in Fig. 4.7.

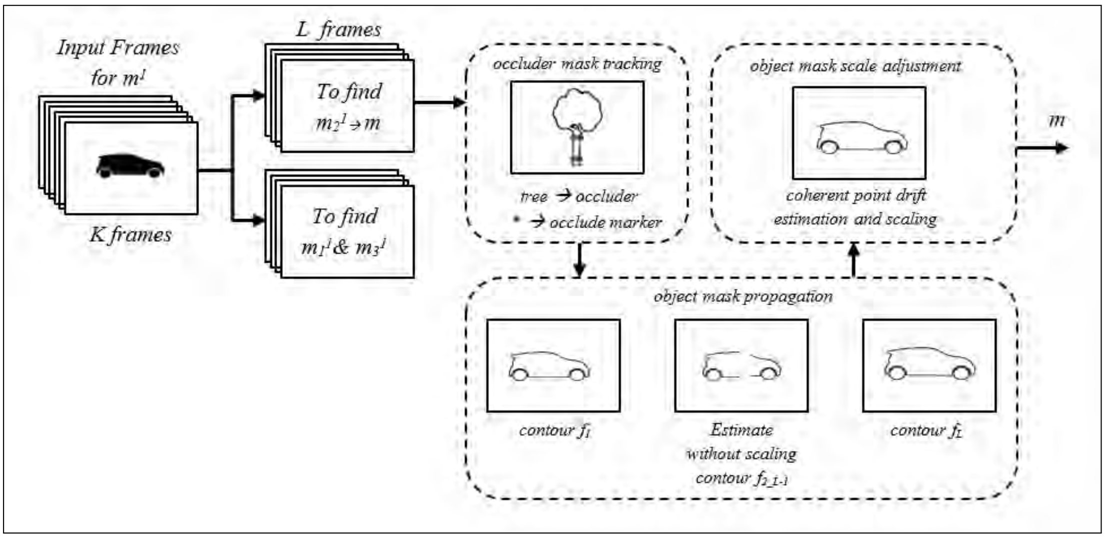


Figure 4.7: Block diagram of the motion-region annotation algorithm of one moving object spanning over K frames. The motion annotation subproblem of type-2 to estimate m spanning over L frames is processed using the proposed algorithm.

4.3.1.1 Occluder mask tracking (m^{occ})

In a sub-problem with distortion, given a set of L frames and the occluder shape marker points, $\mathbf{P}^{occ}_{f_1}$ in the frame f_1 as inputs, the objective of *occluder mask tracking* was to perform shape tracking of the occluder mask in all the remaining $L-1$ frames. Here, the set of frames L in the sub-problem is a subset of the total number of frames N, hence, $L \subseteq K$. The shape marker points of the occluder(s), $\mathbf{P}^{occ}_{f_1}$, in the f_1 frame of the set L were marked by the user through an interactive graphical user interface .

By taking the shape marker points, $\mathbf{P}^{occ}_{f_1}$, of the rigid occluder in the first frame as input, shape tracking of these markers was performed in the rest of the $L-1$ frames. With respect to this shape marker, the occluder mask in the first frame, $m^{occ}_{f_1}$, is given as

$$m^{occ}_{f_1} = region(countour(\mathbf{P}^{occ}_{f_1})) \tag{4.3}$$

while the complete occluder mask set is given as,

$$m^{occ} = \{m^{occ}_{f_1}, m^{occ}_{f_2}, ..., m^{occ}_{f_{L-1}}, m^{occ}_{f_L}\} \tag{4.4}$$

A user is required to define a set of markers (points) around the occluder such that they encapsulate the shape of the occluder. Subsequently, robust SURF features [47] inside the occluder mask, of this n^{th} frame, were estimated as, $F_n = SURF(m^{occ}_{f_n})$. After feature extraction, a point tracker was initialized on the user defined occluder shape markers to estimate their probable position, in the following, $(n+1)^{th}$, frame. Given as,

$$\mathbf{P}^{occ}_{f_n - f_{n+1}} = PointTrackerEst(\mathbf{P}^{occ}_{f_n}) \tag{4.5}$$

The point tracker estimate in the $(n+1)^{th}$ frame was expanded on all sides by an expansion factor λ. The objective was to make sure that even in the case of wrongful tracking by the point tracker, the occluder must be inside the expanded mask. Surf features were again extracted in the λ-expanded mask.

$$F_{n+1} = SURF(region(countour(\lambda \mathbf{P}^{occ}_{f_n - f_{n+1}}))) \tag{4.6}$$

The features F_n and F_{n+1} were matched to yield feature pairs, which were then used to compute a similarity transform.

$$T_s = SimilarityTransform(FeatureMatching(F_n, F_{n+1})) \tag{4.7}$$

This similarity transform, T_s, multiplied with the input shape markers, $\mathbf{P}_{occ_{f_n}}$ results in the shape markers in the next frame.

$$\mathbf{P}^{occ}_{f_{n+1}} = T_s * \mathbf{P}^{occ}_{f_n} \tag{4.8}$$

Using eq. 4.3 for all n, the occluder mask for all the $L-1$ frames of a type-2 sub-problem set with distortions, m^{occ}, can be estimated.

4.3.1.2 Object mask propagation (m^{ini})

Given the object mask in the first frame $m^{ini}_{f_1}$ and the last frame $m^{ini}_{f_L}$ of a subproblem set, the object mask propagation objective was to determine m^{ini}, where

$$m^{ini} = \{m^{ini}_{f_1}, m^{ini}_{f_1}, ..., m^{ini}_{f_{L-1}}, m^{ini}_{f_L}\} \tag{4.9}$$

The user defined input masks are formed independent of occluder to save user time and effort. The output label set, which results when the first frame object mask is propagated forward till the last frame is m^{ini}. This estimate can be utilized to perform non-linear object scale adjustment in the subsequent step.

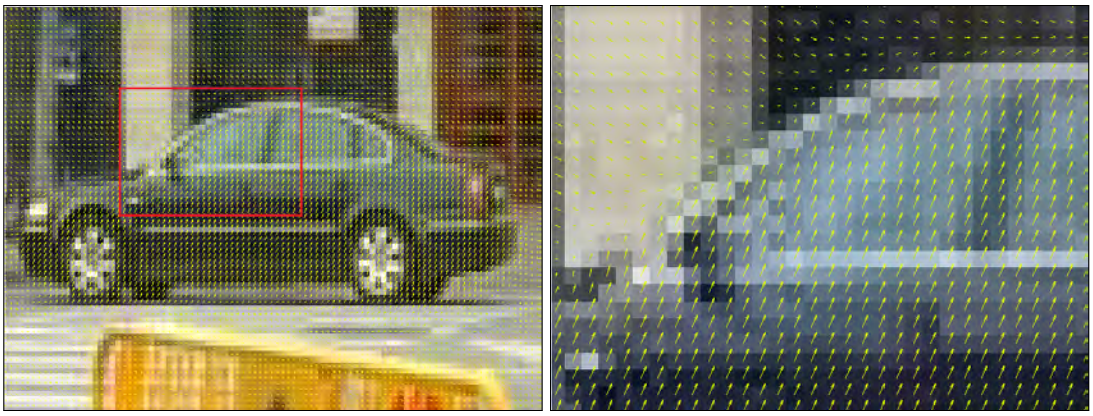


Figure 4.8: **Left:** LDOF vectors overlay on the first frame of a moving object. The direction of flow vectors on the moving object is different from that of the background. More visible in the zoomed image on the 'Right'. **Right:** A zoomed image of the red bounding box from the 'Left' image. Optical flow vectors maintain consistent direction inside the car, but around the object motion boundary and on the background, the vector directions are different.

As a first step for label propagation the forward optical flow, by using the state-of-the-art LDOF [46], was calculated. LDOF supports the estimation of dense optical flow field by integrating rich descriptors into the variational optical flow setting. In [46], the optical flow $\mathbf{w} := (u, v)^T$, is calculated with a comprehensive energy minimization term. These computed flow vectors give an estimate as to where each pixel moved in the following frame.

The given input, $m^{ini}_{f_1}$, contained the labeled pixels pertaining to the moving object region in the first frame. As for every frame n the occluder mask $m^{ini}_{f_n}$ is known, then, for all n, $m^{ini}_{f_n}$ can be updated as,

$$m^{ini}_{f_n} = m^{ini}_{f_n} - (m^{ini}_{f_n} \cap m^{occ}_{f_n}) \tag{4.10}$$

For f_1, it becomes $m^{ini}_{f_1} = m^{ini}_{f_1} - (m^{ini}_{f_1} \cap m^{occ}_{f_1})$. Following this occluder mask subtraction update in the object mask, a set of forward flow vectors of all the pixels in $m^{ini}_{f_1}$ were segregated. In effect, this set contained the pixel-movement estimated by LDOF for all the pixels in the object region. It can be seen from Fig. 4.8 that though the vector directions are robustly detected inside the homogeneous region of the moving object, the estimates around the object boundary are adrift. Hence, as an initial estimate, instead of taking the flow vector per pixel, a 10-bin histogram of vector orientations was computed. All the vectors in the bin with the maximum vector count were separated. The average, direction and magnitude, of this vector set was taken to be the direction and magnitude of the object motion vector, $\hat{w_n}$. In other words, with respect to forward flow, $\hat{w_n}$ is the direction and amount of motion the object mask underwent to reach its new position in the following frame. Formally, if

$$\hat{w_n} = \overline{w_n(\max_{w}(hist(w_n))} \tag{4.11}$$

where $\hat{w_n}$ is the direction vector, then any n^{th} frame in the set of frames L, gives an estimate of the mask position in the following frame by,

$$m^{ini}_{f_{n+1}} = m^{ini}_{f_n} + \hat{w_n} \tag{4.12}$$

By progressively estimating all the frames in the forward direction, m^{ini} was computed.

4.3.1.3 Object mask scaling (m)

Given the object mask in the first frame m_{f_1} and the last frame m_{f_L} of a subproblem set, with m^{ini} already computed, the object mask scale adjustment objective was to determine the final m, where

$$m = (m_{f_1}, m_{f_2}, ..., m_{f_{L-1}}, m_{f_L}) \tag{4.13}$$

Here, it should be noted that $m_{f_1} = m^{ini}_{f_1}$ and $m_{f_L} = m^{ini}_{f_L}$. Hence, the task is to determine m in the remaining $L-2$ frames, from m_{f_2} till $m_{f_{L-1}}$.

A moving object, while in motion inside the FoV, might exhibit a considerable change in depth, perspective and in size. The contour encapsulating a moving object in the first frame might increase or decrease drastically in size and shape in the last frame. An important detail for object mask scale adjustment was to estimate the correspondence of each point on the object contour in the first frame with each point on the object contour in the last frame. As one-to-one correspondence was not possible, there were two options. One was to add or decrease points along the contour from the first frame until the last. This method can result in inaccuracies at each step resulting in error accumulation. Second one was to find a registration between object contours. For this purpose, the point set registration method presented in [48, 49], defined by a function g, was used here. A coherent point drift (CPD) of all the points on the contour in the first frame with reference to the contour in the last frame, was estimated. A 'non-rigid' point drift estimation option was selected, as in some cases perspective changes result in self-occlusion by the object. In this case, the rigidity constraint fails to register the two contours correctly. Hence,

$$m_{f_L}^{CPD} = g(contour(m_{f_1}), contour(m_{f_L})) \tag{4.14}$$

As we get m_{f_1} and $m_{f_L}^{CPD}$ in the same estimated reference, the difference between the two contours was computed to estimate the *linear shape adaptation*, defined as:

$$\kappa = (m_{f_L}^{CPD} - m_{f_1})/L \tag{4.15}$$

Using κ, the scale of all the $L - 2$ frames in the subproblem set can be adjusted, for all n by,

$$m_{f_n} = \kappa(L - n + 1) * m_{f_n}^{ini} \tag{4.16}$$

This adjustment yields the final output m, which gives a shape estimate for the moving object, subtracting the occluder mask, on all the L frames in the sub-problem set. An overall flow of the algorithm is given in Algorithm 1.

This final output m is essentially the annotation mask of the object estimated for the subproblem with distortion (type 2), where [45] failed. Hence, our proposal along with the existing methodology in [45] gives forth a framework, where any object can be annotated semi-automatically with minimum user intervention. Moreover, the given proposal is able to provide an estimated ground-truth annotation in all the frames in the presence of occlusion, change in scale and perspective distortion.

All the algorithm resources including the sub-problem sequences, evaluation source codes, results and the related documentation are publicly available at `http://dixie.udg.edu/anntool/`.

4.3.2 Experimental metrics

Firstly, this section presents the evaluation methods and experimental setup used to assess the motion-region annotation result. Afterwards, the performance of our proposal is exhaustively evaluated, showing both quantitative and qualitative results.

4.3.2.1 Evaluation method

The choice of evaluation criteria is such that a critical insight into the performance of the algorithm can be extracted. There are two factors at play in the motion region-annotation performance assessment: spatial and temporal. So, the goal of the criteria is

Algorithm 1: Motion-Region Annotation Across Occluders

```
Inputs:   Object: $frames \rightarrow \{f_1, ..., f_L\}, m_{f_1}, m_{f_L}$
          Occluder: $P_{f_1}^{occ}$
Outputs: $m = \{m_{f_1}, m_{f_2}, ..., m_{f_L}\}$

function $m^{occ}$ = Occluder mask tracking$(f_1, ..., f_L, P_{f_1}^{occ})$
    where $m^{occ} = \{m_{f_1}^{occ}, m_{f_2}^{occ}, ..., m_{f_{L-1}}^{occ}, m_{f_L}^{occ}\}$
    $m_{f_{1_{occ}}}^{x-d} = region(countour(\mathbf{P}_{occ_{f_1}}))$
    for $n = 1 : L - 1$
        $P_{f_n - f_{n+1}}^{occ} = PointTrackerEst(\mathbf{P}_{f_n}^{occ})$
        $F_{n+1} = SURF(region(countour(\lambda \mathbf{P}_{f_n - f_{n+1}}^{occ})))$
        $T_s = SimilarityTransform(FeatureMatching(F_n, F_{n+1}))$
        $P_{f_{n+1}}^{occ} = T_s * \mathbf{P}_{f_n}^{occ}$
        $m^{occ} = region(countour(\mathbf{P}_{f_n}^{occ}))$
end function

function $m^{ini}$ = Object mask propagation$(f_1, ..., f_L, m_{f_1}, m_{f_L}, m^{occ})$
    where $m^{ini} = \{m_{f_1}^{ini}, m_{f_1}^{ini}, ..., m_{f_{L-1}}^{ini}, m_{f_L}^{ini}\}$
    $m_{f_1}^{ini} \leftarrow m_{f_1}$ ; $m_{f_L}^{ini} \leftarrow m_{f_L}$;
    for $n = 2 : L - 1$
        $m_{f_n}^{ini} = m_{f_n}^{ini} - (m_{f_n}^{ini} \cap m_{f_n}^{occ})$
        $w_n = LDOF(f_n, f_{n+1})$
        $\hat{w_n} = \overline{w_n(\max_w(hist(w_n))}$
        $m_{f_{n+1}}^{ini} = m_{f_n}^{ini} + \hat{w_n}$
end function

function $m$ = Object mask scale adjustment$(f_1, ..., f_L, m^{ini})$
    where $m = (m_{f_1}, m_{f_2}, ..., m_{f_{L-1}}, m_{f_L})$
    $m_{f_1} \leftarrow m_{f_1}^{ini}$; $m_{f_L} \leftarrow m_{f_L}^{ini}$;
    $m_{f_L}^{CPD} = g(contour(m_{f_1}), contour(m_{f_L}))$
    $\kappa = (m_{f_L}^{CPD} - m_{f_1})/L$
    for $n = 2 : L - 1$
        $m_{f_n} = \kappa(L - n + 1) * m_{f_n}^{ini}$
end function

```

to determine how accurately was the annotation propagated in terms of spatial precision as well as temporal evolution.

Spatially, the annotated region in each frame is compared with its respective ground-truth to compute the segmented region overlap performance. This, when accumulated overtime for all frames, gives an average measure of performance. This spatial performance commonly adjudged by *F-score* and *Dice*, which are actually equivalents of each other. Given the evaluation measures true positives (TP), false positives (FP) and false negatives (FN),

$$Sensitivity, S = \frac{TP}{(TP + FN)} \tag{4.17}$$

and

$$Precision, P = \frac{TP}{(TP + FP)} \tag{4.18}$$

are estimated. With S and P known, we compute the F-score F, using the Hungarian method as in [39],

$$F = \frac{2 * S * P}{S + P} \tag{4.19}$$

Developing that expression is equivalent to:

$$D = \frac{2 * TP}{2 * TP + FP + FN} \tag{4.20}$$

F(D) per frame and their average over the set of frames gives a good estimate on how well the resultant annotated region aligns with the reference. The variation in alignment over time and its reasons are however not addressed by these metrics. The values range from 0 to 1. With 0 being worst annotation and 1 meaning that the motion-region coincides perfectly with the ground-truth.

The temporal insight on the evolution of motion-region annotation per frame is grasped profoundly by three more measures, Annotated-Reference region overlap ratio, Occluder-Object size ratio and the change in Hausdorff distance between the reference and annotated regions per frame over time.

The **Annotated-Reference region overlap ratio**, r_n^{a-r} is given by

$$r_n^{a-r} = \frac{m_{f_n}}{m_{GT_n}} \qquad (\forall n = 1, 2, ...L) \tag{4.21}$$

where m_{f_n} and m_{GT_n} are the annotated and reference motion-regions per frame, respectively. This ratio gives an insight on how well the annotated region captures the true ground-truth in terms of its size, its evolution in time exhibits the capability of the algorithm to cope with the ground-truth even if the annotation is corrupted in the middle frames. Its value varies between 0 and 1, with 0 indicating no overlap and 1 indicating complete overlap of the two masks.

The **Occluder-Object size ratio**, r_n^{c-b} is given by

$$r_n^{c-b} = \frac{pixels(m_{f_n}^{occ}) \cap pixels(m_{f_n})}{pixels(m_{f_n})} \qquad (\forall n = 1, 2, ...L) \tag{4.22}$$

where $m_{f_n}^{occ}$ is the occluder region in each frame. This is the ratio of the overlapped area of the occluder and annotated regions, with the total annotated motion region. This measure gives an idea on how much of the motion region is occluded by the occluder. The algorithm's performance in these regions, over time, tells us how robust the algorithm is to the size of the occluder. Here the r_n^{c-b} value varies between 0 and 1, with 0 indicating no occlusion and 1 indicating that the object is complete occluded by the occluder. It should be noted that if r_n^{c-b} becomes 1, the algorithm will fail, as it requires some part of the moving object to be visible at all times.

The **Hausdorff distance** H^{dist} between the reference and annotated regions is given as,

$$H_n^{dist} = HausDist(m_{f_n}, m_{GT_n}) \qquad (\forall n = 1, 2, ...L) \tag{4.23}$$

where $HausDist$ indicates the hausdorff distance implementation function. Intuitively, H^{dist} finds the point p from the set m_{f_n} that is farthest from any point in m_{GT_n} and measures the distance from p to its nearest neighbor in m_{GT_n}. This measure gives an insight as to how far off the worst annotated motion-region point is with respect to the ground-truth. If evaluated over time, it gives an idea of the temporal robustness as well as the reliability of the algorithm. Here, a good annotation means that the H^{dist} value is close to zero, given in pixels. A greater H^{dist} value would indicate the magnitude of misalignment of the annotated mask with the reference.

4.3.2.2 Experimental setup

The performance of the motion-region annotation algorithm was tested on a newly formed sub-problem dataset. This was done by taking a total of 25 snippets from the region-based long sequences benchmark explained in Section 3.3.1. These 25 video sequences are sub-problems of annotation, when the moving object underwent occlusion. Among the 25 sets of frames, 20 contain static occluders, and the remaining 5 contain moving occluders. A few examples of this motion-region annotation dataset are shown in Fig. 4.9.

The 20 sequences with static occluders encompass 15 with one occluder and 5 with two occluders, as listed in Table. 4.2. The depth of each moving object being annotated is also indexed in three categories, low, medium and high. A moving object at low depth means that the object and occluder are near the camera, so they appear big in size and may have distinct features contained in them. A high depth means that the object size is small in the field of view. In this case, the occluder might be big or small, depending upon its own depth.

In addition to the 20 sequences with static occluders, 5 more sequences were taken with moving occluder. In this case, the occluder mask is already given, as these moving occluders are motion-regions of the same sequence, which have already been annotated. The moving object depth in these sequences is also listed. All these sets of frames contain a single moving occluder.

To establish the efficacy of our work, we evaluate the performance of our algorithm in comparison with other state-of-the-art contributions. The choice of methods to utilize is limited due to a number of factors, namely, availability of code, applicability on the proposed scenario (ability to propagate the label across occluders and be able to recover the shape of the moving object) and computational time. Of the listed factors, applicability of the algorithms in our scenario is a limiting factor as most algorithms fail, when motion label is propagated across an occluder. There are tracking algorithms, which are able to perform this task but they provide bounding boxes on the moving object instead of moving object boundary. Hence, we present a comparative analysis with two recent methods, a probabilistic method [17] and a learning-based method [18]. Both are moving object segmentation methods, which give the moving object motion boundary as the output. These methods do not start with known initial object boundary as in our method, so to make it fair to them, we consider there results

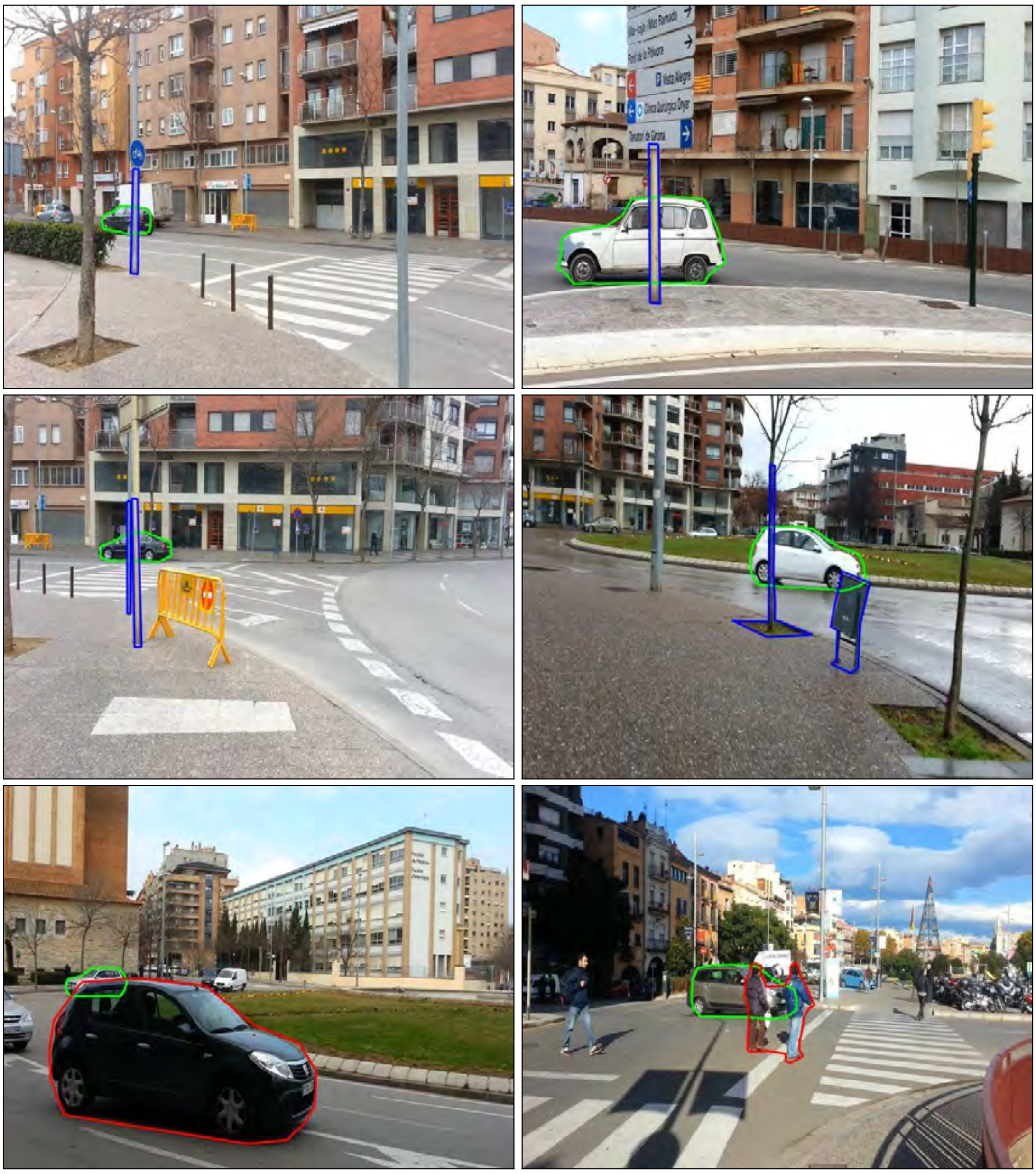


Figure 4.9: The moving objects being annotated in the given examples are captured by a green contour around them. The static occluders are shown in blue, while the moving occluders are shown in red, bounding regions around them, respectively. **Top Row:** Two examples of moving objects going across single static occluder. The black car in the left image has high depth, whereas the white car in the right image has low depth, near the camera, **Middle Row:** Two examples of moving objects going across two static occluders. The left image is high depth and the right image is medium depth, **Bottom row:** Two examples of moving objects going across moving occluders. The left image has very high depth and the right image has medium depth. The moving occluder in the left image is the black moving car, which occludes our desired moving object almost completely. In the right image the moving car is occluded by moving people.

Table 4.2: A summary of the features of the motion-region annotation dataset. Acronyms are **Avg.:** Average. In object depth, **Lw:** Low, **Md:** Medium, **Hg:** High.

Motion-Region annotation dataset features				
Datasets	**Dataset Features**			
	Total sequences	Total frames	Avg. frames	Object Depth
Static Occluder (One)	15	340	22.7	Lw/Md/Hg
Static Occluders (Two)	5	166	33.2	Md/Hg
Moving Occluders	5	177	35.4	Md/Hg

correct on any motion they were able to correctly segment around the ground-truth. This consideration gives an advantage to the algorithms in terms of motion estimation on or around the moving object, but in effect makes them not applicable for moving object occluder sequences. Furthermore, as these methods do not estimate occluder boundary separately, hence the *Occluder-Object size ratio*, r_n^{c-b} is not calculated for them.

A 64-bit Intel i7 core 3.4 GHz machine with 16GB RAM was used for processing, except LDOF calculation, which was run in a similar server machine with 128GB RAM. All the scripts and results related to the experiments done are publicly available online with the motion-region dataset. Also, the scripts are designed as such to be able to incorporate any new algorithm for standardized comparison of results.

4.3.3 Experimental results

The experimental results are both quantitatively and qualitatively analyzed. The details of each are elaborated in the following sections.

4.3.3.1 Quantitative results

The results of the motion-region annotation algorithm on the presented dataset are given in Table 4.3, 4.4 and 4.5. The accumulative average F-score on static occluders as well as moving occluders reaches up to 95%.

Upon static object occlusion, a maximum F-score of 98% is achieved for *seq-03*, and the lowest is 73% for *seq-17*, where, on average, 21% motion-region area was occluded by 2 occluders, as given by the corresponding r^{c-b}. For moving occluders, a maximum F-score of 97% is achieved for *seq-24*, even in the presence of 58% occlusion. The lowest F-score in moving occluders is 94%, which is achieved even when, on average, 68% of the motion region was occluded. Observing the results in Table 4.4, it is observed that the

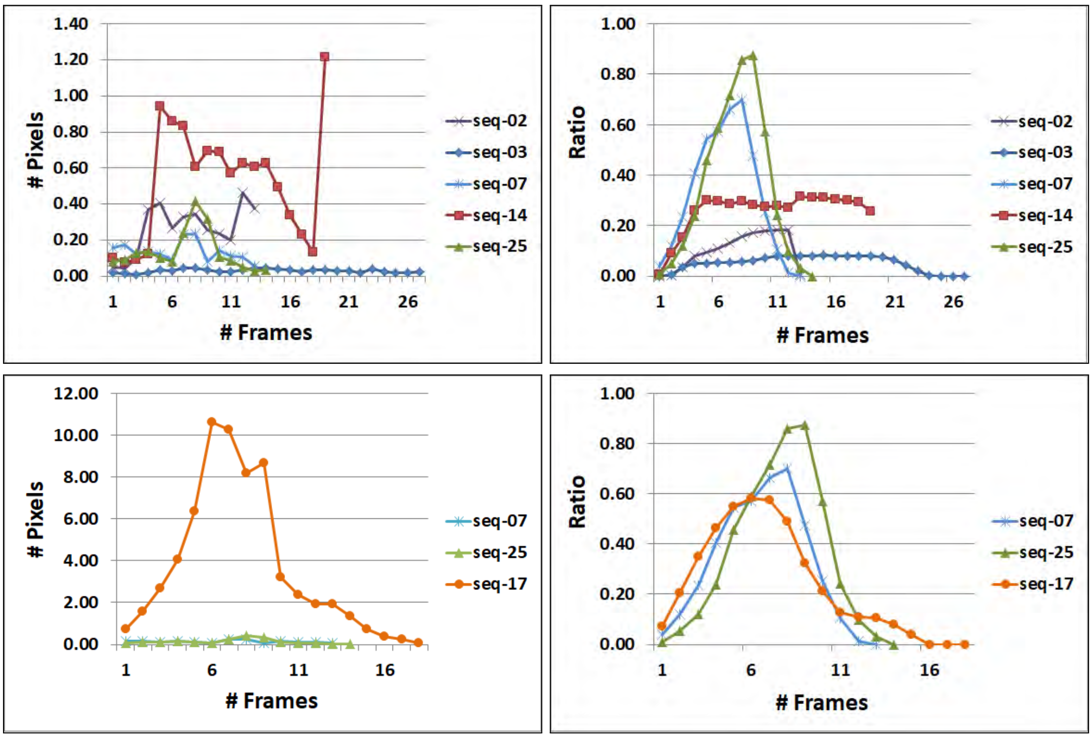


Figure 4.10: **Top:** The temporal evolution of performance measures of five sequences, *seq-02*, *seq-03*, *seq-07*, *seq-14* and *seq-25*. **Bottom:** The temporal evolution of performance measures of three sequences, *seq-07*, *seq-25* and *seq-17*. **Left:** A visualization of the change in H^{dist} over time in each frame (in pixels). **Right:** The change in occluder-object ratio r^{c-b} over time in each frame (ratio).

overall performance of the two Probabilistic [17] and Learning-based [18] algorithms is not suitable to be used as ground-truth. In general, these algorithms do an acceptable motion segregation when the motion is small and the moving object depth in the scene is relatively small. The algorithms fail when the object is too large or too small.

The occluder-object overlap ratio, r^{c-b}, indicates the percentage amount of annotated motion-region being occluded. A higher value of this ratio signifies that the most part of the moving object is covered. It can be seen from the results that even with high r^{c-b}, the algorithm is able to propagate the label correctly in the following frames. In sequences *seq-07, seq-19, seq-22, seq-24, seq-25*, where the occlusion percentage reaches up to 44%, 38%, 38%, 58% and 88% respectively, the algorithm performs as high as 97% and never goes below 86%.

The annotation-reference overlap ratio r^{a-r} and Hausdorff distance H^{dist} should be understood in conjunction. r^{a-r} gives a measure of how much of the propagated annotation conforms correctly with the ground-truth, while H^{dist} measures how far the worst propagated label is from the ground-truth annotation. Here, with static

Table 4.3: A summary of the results of the label propagation algorithm on the motion-annotation dataset featuring static occluders. Acronyms are **Seq.:** Sequences, **S.:** Sensitivity, **P:** Precision, **F:** F-score, **D:** Dice score.

Results on Motion-Region Annotation Dataset having Static Occluders							
Seq. attributes		**Spatial importance**			**Temporal importance**		
Name	Frames	S	P	F(D)	r_n^{a-r}	r_n^{c-b}	H^{dist}
One static occluder							
seq01	24	0.94	0.96	0.95	0.90	0.01	0.29
seq02	15	0.87	0.97	0.92	0.84	0.09	0.26
seq03	29	0.98	0.98	0.98	0.96	0.05	0.03
seq04	21	0.99	0.95	0.97	0.94	0.06	0.12
seq05	15	0.99	0.97	0.98	0.95	0.03	0.07
seq06	20	0.97	0.95	0.96	0.92	0.04	0.10
seq07	15	0.96	0.95	0.95	0.91	0.44	0.14
seq08	14	0.98	0.94	0.96	0.92	0.12	0.12
seq09	15	0.99	0.91	0.95	0.91	0.07	0.26
seq10	73	0.93	0.93	0.93	0.87	0.07	0.14
seq11	20	0.96	0.95	0.95	0.91	0.21	0.10
seq12	19	0.93	0.96	0.95	0.90	0.17	0.15
seq13	18	0.95	0.96	0.96	0.91	0.11	0.11
seq14	21	0.92	0.88	0.90	0.82	0.26	0.52
seq15	21	0.98	0.89	0.94	0.88	0.09	0.45
Two static occluders							
seq16	25	0.93	0.96	0.94	0.89	0.11	0.19
seq17	20	0.72	0.73	0.73	0.57	0.21	3.64
seq18	62	0.98	0.91	0.94	0.89	0.03	0.40
seq19	38	0.91	0.81	0.86	0.75	0.38	0.48
seq20	21	0.97	0.95	0.96	0.93	0.08	0.10
Overall cumulative results with static occluders							
Average	25.3	0.96	0.93	0.95	0.90	0.08	0.38
Max	73	0.99	0.98	0.98	0.96	0.44	3.64
Min	14	0.72	0.73	0.73	0.57	0.01	0.03

Table 4.4: A summary of the comparative results of metrics on the motion-annotation dataset featuring static occluders. Acronyms are **Seq.:** Sequences, **F:** F-score, **D:** Dice score, r_n^{a-r}: Annotated-Reference region overlap ratio and H^{dist}: Hausdorff distance.

Results on Motion-Region Annotation Dataset having Static Occluders									
	Probabilistic [17]			**Learning [18]**			**Ours**		
Name	F(D)	r_n^{a-r}	H^{dist}	F(D)	r_n^{a-r}	H^{dist}	F(D)	r_n^{a-r}	H^{dist}
One static occluder									
seq01	0.78	0.64	1.35	0.79	0.65	1.29	0.94	0.89	0.19
seq02	0.74	0.59	2.96	0.74	0.59	2.17	0.73	0.57	3.64
seq03	0.88	0.79	1.17	0.82	0.69	2.10	0.95	0.90	0.29
seq04	0.82	0.69	1.21	0.78	0.64	1.23	0.92	0.84	0.26
seq05	0.87	0.77	1.56	0.81	0.68	2.31	0.98	0.96	0.03
seq06	0.45	0.29	23.33	0.85	0.73	1.46	0.97	0.94	0.12
seq07	0.00	0.00	—	0.84	0.72	3.00	0.98	0.95	0.07
seq08	0.41	0.26	13.84	0.83	0.72	1.54	0.96	0.92	0.10
seq09	0.70	0.54	4.84	0.81	0.69	1.81	0.94	0.89	0.40
seq10	0.63	0.45	3.80	0.75	0.60	2.43	0.95	0.91	0.14
seq11	0.51	0.34	4.94	0.53	0.36	7.32	0.86	0.75	0.48
seq12	0.79	0.65	3.86	0.73	0.58	3.31	0.96	0.92	0.12
seq13	0.74	0.58	6.56	0.78	0.64	3.73	0.95	0.91	0.26
seq14	0.05	0.02	24.37	0.38	0.23	13.53	0.93	0.87	0.14
seq15	0.48	0.32	7.05	0.44	0.28	10.86	0.95	0.91	0.10
Two static occluders									
seq16	0.68	0.52	4.33	0.80	0.67	1.79	0.95	0.90	0.15
seq17	0.85	0.74	0.84	0.77	0.62	2.03	0.96	0.91	0.11
seq18	0.80	0.67	1.88	0.77	0.62	2.67	0.96	0.93	0.10
seq19	0.61	0.44	5.84	0.64	0.47	3.78	0.90	0.82	0.52
seq20	0.75	0.60	5.60	0.71	0.56	7.36	0.94	0.88	0.45
Overall cumulative results with static occluders									
Average	0.63	0.50	6.28	0.73	0.59	3.79	0.93	0.88	0.38
Max	0.88	0.79	24.37	0.85	0.73	13.53	0.98	0.96	3.64
Min	0.00	0.00	0.84	0.38	0.23	1.23	0.73	0.57	0.03

Table 4.5: A summary of the results of the label propagation algorithm on the motion-annotation dataset featuring moving occluders. Acronyms are **Seq.:** Sequences, **S.:** Sensitivity, **P:** Precision, **F:** F-score, **D:** Dice score.

Results on Motion-Region annotation dataset having moving occluders							
Seq. attributes		**Spatial importance**			**Temporal importance**		
Name	Frames	S	P	F(D)	r_n^{a-r}	r_n^{c-b}	H^{dist}
seq21	42	1.00	0.92	0.96	0.92	0.12	0.14
seq22	25	0.94	0.98	0.96	0.93	0.38	0.08
seq23	58	0.94	0.94	0.94	0.89	0.29	0.33
seq24	36	0.95	0.99	0.97	0.94	0.58	0.08
seq25	16	0.94	0.95	0.94	0.89	0.68	0.14
Overall cumulative results with moving occluder							
Average	33.2	0.95	0.94	0.95	0.90	0.30	0.16
Max	58	1.00	0.99	0.97	0.94	0.68	0.33
Min	16	0.94	0.92	0.94	0.89	0.12	0.08

occluders, it can be seen that maximum r^{a-r} of 96% is achieved in *seq-03* with H^{dist} as low as 0.03 pixels on average. The lowest overlap of 57% is experienced in *seq-17* with H^{dist} as high as 3.64 pixels on average. It is interesting to note that these results are consistent with the performance exhibited by F-score.

One thing which cannot be appreciated through these average performance measures is the capability of the algorithm to recover, in case of failure in the intermediate frames. A temporal evaluation per frame gives a better insight on this behavior. This temporal evaluation is shown in Fig. 4.10, where the evolution of H^{dist} and r^{c-b} of some selected sequences per frame can be visualized.

In Fig. 4.10, top row, the temporal progress of H^{dist} and r^{c-b} in sub-problem sets of frames from five video sequences are shown. It can be seen that in *seq-02* and *seq-03* as the percentage occlusion of the object, r^{c-b}, remains below 20%, then the farthest point of the annotated labeled contour from the reference label contour, H^{dist}, never increases more than 0.3 pixel. As r^{c-b} increases to almost 32% in *seq-14*, the maximum propagation error in terms of distance stays within 1.5 pixel distance. It can also be appreciated that in *seq-07* where even with a 70% peak r^{c-b}, the H^{dist} never goes beyond 0.25 pixels. This trend is also observed in the moving occluder sequence *seq-25*, where even in the presence of 88% peak r^{c-b}, the annotation error in term of H^{dist}

remains with in 0.5 pixels for all frames. In general, the algorithm performs well in all the sequences even in the presence of high percentage of occlusion of the moving object. While in Fig. 4.10, bottom row, the temporal progress of the same measures in three video sequences are shown. It can be seen that only, *seq-17* behaves differently. Here, in the presence of 60% occlusion, which is less than that of *seq-07* and *seq-25*, the maximum H^{dist} goes up to 11 pixels.

Another perspective of evaluation is to observe the performance of the algorithm on relatively long set of sequences. Taking one from each type, we see that in *seq-10*, *seq-18* and *seq-23* with 73, 62 and 58 frames, respectively, the algorithm had an average F-score of 95% and an average H^{dist} of 0.32 pixels. These sequences exhibit a variety of characteristics, where the moving object is, at a high depth in *seq-10*, at a medium depth in *seq-18* going across two occluders and at a medium depth in *seq-23* going across moving non-rigid occluders. The average performance shows that the algorithm is not affected by the length of frames as much as the type of motion in them.

4.3.3.2 Qualitative results

The qualitative results give a visual and intuitive evaluation of the algorithm. In Fig. 4.11 and 4.12, the results of motion label propagation are shown, with one occluder, two occluders and moving occluder.

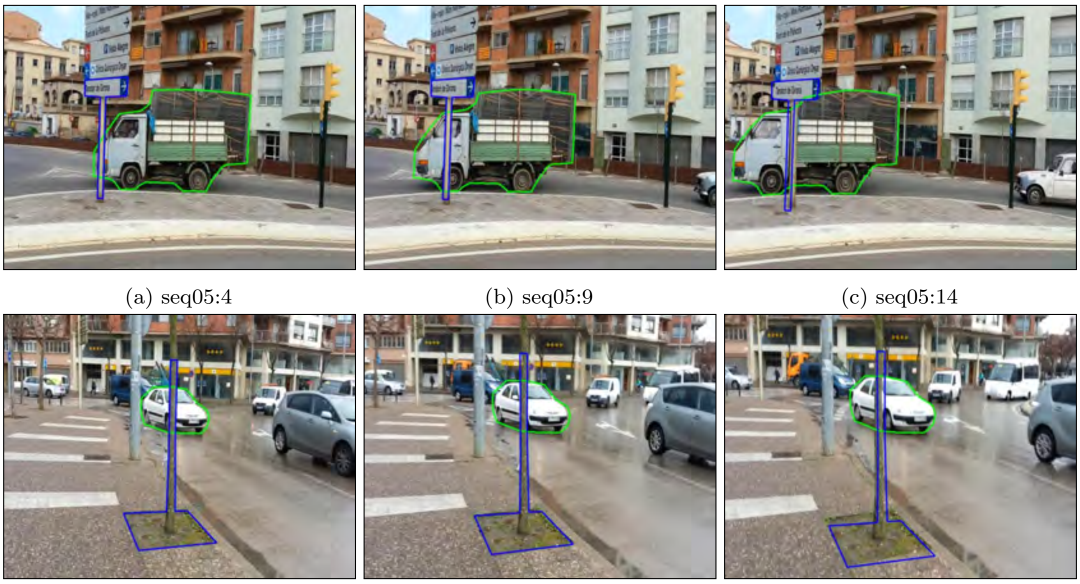
In Fig. 4.11, three different frames, first, middle and last, of two sequences with a single static occluder are shown. In the top row from *seq-05*, a large truck is seen going across a direction post. The truck has a low depth in the field of view, meaning it is close to the camera. The average occlusion percentage is 3%, but the issue to note is that the whole body of the moving object undergoes occlusion at least once during the complete motion. The occluder mask was created with a few clicks around the direction post and it was tracked as mentioned in Section. 4.3.1.1 the occluder mask is robustly tracked. This robust result facilitates the shape propagation of the motion mask across all frames. As the shape and perspective change of the moving object is minimal, the results achieved are as good as 98%.

In the bottom row of the figure, three frames of the sequence *seq-08* are shown. The white car undergoes an occlusion by a tree stem. The car moves across multiple frames coming towards the camera, which changes its depth. This can be verified from the first and the last frame, as the size variation of the car is visually apparent. The thin tree stem occluder is marked in the first frame by defining a few points around it. Here, it can be seen that the area around the trunk is also marked. As the tree trunk is quite thin, the soil area around the trunk reinforces the SURF feature extraction and matching, resulting in a better tracked occluder. The linear change adaptation factor κ, as explained in Section. 4.3.1.3, gives a good estimate of the change in depth of the car in each progressive frame. So even in the case of depth change, the achieved F-score is 96%.

In the top row of Fig. 4.12, three frames from *seq-19*, where the moving object is occluded by two occluders, are shown. It can be seen that the white car gets occluded by a lamp post and a thin tree trunk. Over the course of the motion, the size of the moving object changes considerably as it moves towards the camera. The occluder masks are marked in the first frame of the sequence, and it can be seen that the masks are well tracked even until the end. The object starts moving from a high depth and comes towards the camera to medium depth. With such a big change in depth, and even with 38% occlusion on average, a F-score of 86% is achieved. Here it can be appreciated that the algorithm possesses the capability to map a small contour in the starting frames to an expanded large contour in the ending frames with consistency in shape, and vice versa.

In the bottom row of Fig. 4.12, three frames from *seq-25*, where the moving object is occluded by a single moving occluder, are shown. An extreme case is present in this sequence, as the moving object is at a higher depth and has a small size, as compared to the moving occluder, which is at a low depth, hence quite large in size. On average the occlusion ration reaches up to 88%. Even in the presence of such occlusion, due to reliable LDOF calculation, as mentioned in Section 4.3.1.2, our algorithm performs well, achieving 94% F-score. Here, the moving occluder is assumed to have been previously annotated, therefore, the occluder mask marking and tracking is not performed.

In Fig. 4.12, we also show three frames from *seq-17*, where the moving object is occluded by two occluders. It can be seen that the black car goes across two lamp posts. The occluder masks are marked in the first frame and tracked until the last. In the last frame, the tracker losses the shape of a marker but it does not effect the result as there is no overlap between the wrongly tracked occluder mask and the moving object mask. Besides the occluder mask, the motion label propagation as shown by a green contour around the black car failing to propagate the label correctly. The propagated labels move ahead of the ground-truth, this means that the maximal velocity count consensus is making the mask move in the right direction but not with the correct magnitude. Upon further investigation on the obtained results, we observed that there are two competing hypothesis on the magnitude of the motion vector. Here, the wrong hypothesis edges past the correct one with a small difference. This occurs due to



(a) seq05:4

(b) seq05:9

(c) seq05:14

(d) seq08:2

(e) seq08:8

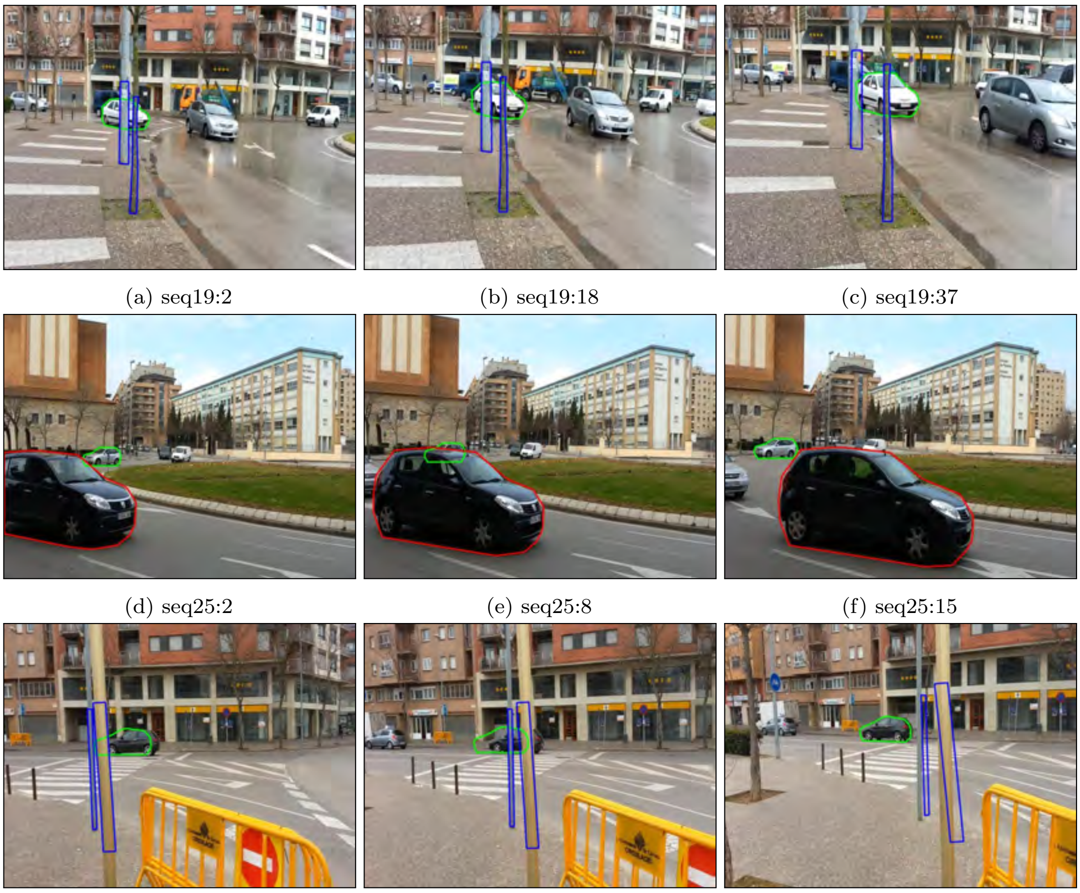
(f) seq08:13

Figure 4.11: Motion annotation result on three frames of two sequences containing single static occluder. The motion and the occluder masks are shown in green and blue contours, respectively. In the sub-figure description code 'seqXX:YY', XX is the sequence number and YY is the frame number. **Top Row:** The subfigures a, b and c, show the frames 4, 9 and 14, respectively, from the sequence *seq-05*. The moving truck is occluded by the static direction post (F-score: 98%). **Bottom Row:** The subfigures d, e and f, show the frame 2, 8 and 13, respectively, from the sequence *seq-08*. The white car is occluded by the static thin tree trunk (F-score: 96%).

the background around the car because the LDOF calculated at the edges of the car gets tampered due to the color similarity between the car and the background. This limitation could be overcome by introducing a factor catering for background similarity in the maximal vector consensus.

4.3.4 Conclusion

In our region annotation proposal, a framework to address the problem of motion annotation in the presence occlusion, depth change and perspective distortion was presented. Our approach in integration with with an existing methodology [45] formulates a framework to overcome the prevailing limitations in motion-region annotation. It was shown that with minimum manual intervention and with best utilization of the expert-user time, the generation of ground-truth label for moving objects can be done even in the presence of real distortions. A three pronged approach was taken, where first the occluder mask was tracked in subsequent windows with SURF feature matching and similarity transformation. Then, the object mask propagation was done by



(a) seq19:2 (b) seq19:18 (c) seq19:37

(d) seq25:2 (e) seq25:8 (f) seq25:15

(g) seq17:3 (h) seq17:10 (i) seq17:19

Figure 4.12: Motion annotation result on three frames of three sequences is shown. The motion mask is shown in green contours. The static occluder masks are shown in blue, while the moving occluder mask is shown in red, contours. In the sub-figure description code 'seqXX:YY', XX is the sequence number and YY is the frame number. **Top Row:** The subfigures a, b and c, show the frames 2, 18 and 37, respectively, from the sequence *seq-19*. The moving white car is occluded by two static occluders, a lamp post and a tree trunk (F-score: 86%). **Middle Row:** The subfigures d, e and f, show the frames 2, 8 and 15, respectively, from the sequence *seq-25*. The gray car is occluded by the moving black car (F-score: 94%). **Bottom Row:** The subfigures g, h and i, show the frames 3, 10 and 19, respectively, from the sequence *seq-17*. The black car is occluded by the two static lamp posts (F-score: 73%).

computing maximal consensus motion vectors from the state-of-the-art LDOF estimation. And finally the scale adjustment of the propagated object mask was performed by first to last frame point-set registration, coupled with linear adaptation factor κ. For evaluation purposes, we also proposed a motion annotation dataset with 25 sequences, containing single and multiple static and moving occluders. We presented a detailed quantitative and qualitative analysis of the methodology to show that it can be reliably used for label propagation in sequences with occlusion and other real noises, reaching an average F-Score as high as 95%. We have also shared the source codes, results and the related documentation publicly for the community to use it and to perform further improvements in this methodology.

4.4 Limitations and open issues

Two motion annotation methodologies were discusses in this chapter. One proposal was a trajectory-based motion annotation platform, where any video sequence, with an apriori tracked trajectory matrix, can be semi-automatically annotated. The second proposed methodology was a motion-region annotation framework, where moving object labels were propagated across occlusions. The propagation was based on occluder mask tracking, object mask propagation and object mask scaling. Though, the methodologies proposed a solution to address the motion annotation problem in their respective domains, there are still limitations and certain open issues that need to be dealt with.

The noisy trajectories in the trajectory-based dataset annotation tool are not comprehensively managed. These are the trajectories, where the tracking failed, and ended up with a trajectory encapsulating two or more motions in it. There can be two types of such trajectories,

- A set of wrongly tracked trajectories, which are partly tracked over moving objects, while part of it tracks the background.
- A set of wrongly tracked trajectories, which captures the motion of multiple moving objects in it, but not the background.

In the current settings, as presented in Section 4.2.1, these trajectories are assigned a label which they are more representative of, as estimated by the expert-user. The result of a MS algorithm on these set of trajectories can be misleading. An MS algorithm classify a trajectory to be the motion, which it was less representative of. In that case, it will be considered a classification error, whereas the algorithm correctly identified one of the two possibilities of motion label assignment. This problem further complicated when considered that two possibilities of noisy trajectories. Some possible solutions might be to either get rid of these trajectories altogether, or break them down into parts or devise a framework where multiple labels can be assigned to a trajectory. All these solutions require the identification of such trajectories. There recognition among the complete cluster of trajectories would take a lot of expert-user time.

In the proposed region annotation proposal, the motion-regions of moving objects were propagated across static and moving occluders. The application of the proposal is only valid for partially occluded objects. Its applicability on completely occluded objects is a limitations. In the present framework, addressing this limitation is difficult. If the subproblem containing complete occlusion is subdivided into part with partial occlusion, then the algorithm can deal with them. In a post processing step, the two subdivided part can be joined to represent one moving object label. The post processing can be utilize the global information related to color, texture, shape and object trajectory in the FoV. These features in a clustering framework can yield improved results.

The pixel accuracy of the propagated object masks and the tracked occluder masks is another limitation. An algorithm which can locally improve the accuracy around occluder-occlusion regions was presented in [85]. The coarse to fine region-based sobolev descent [85] was used in shape tracking. In their proposal, its application was exhibited in the presence of human motion, with a condition that the amount of motion per frame is limited to a change in a few pixels. This methodology can be applied to our per frame propagation output as the difference in original object location and our propagated output will hold as our results show that the difference in the mask and the ground-truth never exceeds a few pixels.

Chapter 5

Conclusions

Motion segmentation is an active research field which at present faces impediment in its progress. The forthcoming algorithms still adopt the same set of assumptions (known and limited number of motions, complete trajectories, no perspective distortions, etc.) and their performance is mostly assessed on less representative datasets. With the aim to overcome these limitations, we provide a collection of diverse challenging datasets comprising long and short sequences of real-life natural scenes. The ground-truth annotation is given in the form of trajectory- and region-based labels on all the motions of all the frames. Such annotation facilitates the evaluation of motion segmentation methods on a single common platform. Moreover, an increased number of motions and frames per sequence, together with the presence of real distortions, provide a new challenge for the community.

To set an initial benchmark, we evaluate the performance of six state-of-the-art algorithms on all the new sequences of the proposed database. Our evaluation metrics and obtained results reveal that the problems of real-life distortion processing, separation of similar motions and label recovery after occlusion still remain a serious challenge. Therefore, we believe that our new database will provide an opportunity for a deeper understanding of the motion segmentation problem, and will push the boundaries of research.

The most critical step in the creation of dataset is the generation of ground-truths. In the case of motion represented by trajectories, these set of trajectories are grouped in clusters with unique labels to distinguish one motion from another. The MS literature review showed that there are many proposals to address the motion segmentation problem, but none have proposed a standardized way of generating the ground-truth on motion trajectories. In our work, we proposed a semi-automatic motion trajectory annotation tool, TAT. The tool was comprehensively tested to provide the trajectory labeling in our trajectory-based mong sequences dataset. We also evaluated the performance of out tool, in several test cases to quantify the behavior of the tool performance.

In the case of motion-region labeling, generally, a semi-automatic approach is taken for the ground-truth annotations, where the user defined object masks in a few frames are propagated with accuracy in the remaining frames. The state-of-the-art contains several techniques to address the region-annotation problem, but their performance in the presence of partial and complete occlusion, perspective distortion, stopping motion, illumination changes, depth change and shadows are inadequate. In our work, we presented a framework to address the problem of motion annotation in the presence of occlusion, depth change and perspective distortion. Our proposal works in integratation with an existing methodology [45] to formulate a framework to overcome the prevailing limitations in the state-of-the-art. Our focus was that with minimum expert-user intervention, the generation of ground-truth label for moving objects should be done even in the presence of real distortions. A three pronged approach was taken where first the occluder mask was tracked in subsequent windows, with SURF feature matching and similarity transformation. Then, the object mask propagation was done by computing maximal consensus motion vectors from the state-of-the-art LDOF estimation. And finally the scale adjustment of the propagated object mask was performed by first to last frame point-set registration couple with linear adaptation factor κ. For evaluation, we also presented a motion annotation dataset with 25 sequences, containing single and multiple static and moving occluders. We presented a detailed quantitative and qualitative analysis of the methodology to show that it can be reliably used for label propagation in sequences with occlusion and other real noises. We have also shared the source codes, results and the related documentation publicly for the community to use it and to perform further improvements. All ur shared data is publicly available at `http://dixie.udg.edu/udgms/`.

5.1 Summary of the thesis

We started our research by developing an understanding of Motion cue in videos, more specifically its usage in the motion segmentation problem. The introduction chapter briefly discussed the problems being faced by the motion segmentation community. In particular, the need of modern datasets with long video sequences was discussed, which became the cornerstone of the main motivation of this work. The factors effecting the current state-of-the-art and the need of new proposals were described.

In continuation, the need to benchmark the datasets with respect to modern algorithms was explained. This became a precursor to move towards motion annotation. The reasoning which led to developing motion annotation tools and the modalities were also argued. We highlighted the fact that modern datasets can only be made representative, if automatic annotation or label propagation research is given due attention for devising new methods.

Chapter 2 extended this study by an in-depth review of the literature addressing a range of motion segmentation related topics. We took a detailed look in to the MS datasets used by the community in their research. We classified the different types of datasets based on their utility and application. Then, we sifted through the literature to see what modern effective MS algorithms are being used these days. Relevant algorithms were distinguished based on their usage in different distortions. We also segregated the techniques, which were applicable on long sequences from the ones, which can address only short sequences. Besides this review, we also looked into the motion annotation problem. A detailed study in understanding the state-of-the-art in motion-trajectory and motion-region annotation was discussed. Based on the conclusions drawn from these reviews, we directed our work towards the development of a collection of motion segmentation benchmark datasets along with motion annotation techniques to create such datasets.

In Chapter 3, we described the making of a collection of motion segmentation datasets. We targeted all the prevailing state-of-the-art modalities of MS datasets, therefore, we proposed trajectory- and region-based, long and short sequences, datasets. The reason behind the making of long and short datasets was discussed in detail. This chapter explains each step of acquisition, tracking and annotation. Further along, the benchmarking of the presented datasets with recent MS algorithms was also presented in detail. The results of the benchmark clearly established the difficulty of the compiled problem. This difficulty level of the datasets provides an opportunity to MS community to present better algorithm so as to address the motion segmentation problem better.

In Chapter 4, the description of motion annotation tools was inscribed. Initially, a comprehensive motion-trajectory labeling tool was presented. The processing pipeline of the tool was discussed in detail, with an elaboration of each block separately. Its application in the making of our trajectory-based long sequences dataset was also given, with an analysis of its performance on different video types. Following the trajectory annotation tool, a comprehensive region annotation framework was also presented. The framework formulation based on the integration with an existing methodology [45], and the associated algorithms to deal with distorted subproblems were explained in detail.

The results of applying the framework on our proposed motion annotation dataset were discussed in detail. Furthermore, the usage in the presence of static and moving occluders was also analyzed, showing satisfactory results on the tests performed.

5.2 Contributions

The following are the major contributions of this thesis:

- A collection of motion segmentation benchmark datasets, with an emphasis on the two modalities prevalent in the state-of-the-art.

  - Trajectory-based long and short sequences datasets: A total of 19 long sequences with 200 motions and 162 short sequences with 442 motion.
  - Region-based long and short sequences datasets: A total of 20 long sequences with 235 motions and 150 short sequences with 440 motion.

  Two more datasets of each modality were also presented containing no missing data. This approach was taken so that the state-of-the-art MS algorithms could be tested.

- The creation of benchmark based on the recent MS algorithms. The OB algorithm applied was used to test the long sequences datasets. Furthermore, the OB algorithm [39] along with LS3C [44], SSC [31,86], ALC [43,87], ELSA [34] and LRR [33] were used to benchmark the short sequences. This is the most recent benchmark with sequences encapsulating such difficulty. It creates room for the MS community to propose algorithms, which can solve the problem more comprehensively. The quantitative and qualitative comparison of proposed algorithms can be compared with this standardized benchmark. All the resources including the datasets, source codes and evaluation scripts are available at `http://dixie.udg.edu/udgms/`.

- A motion-trajectory annotation tool proposal to annotate any trajectory-based sequence. The publicly available tool provided a standardized platform for the annotation of trajectory clusters in video sequences. With two initialization modes, the dataset retains flexibility of application. The tool can be used to reinforce trajectory-based MS datasets with more sequences containing complex motions.

- The motion-region annotation framework to provide labels on motion-regions undergoing distortions. The kind of noise our framework can deal with are occlusions, depth change and perspective distortion. Our proposal performs semi-automatic label propagation across partial occlusions. The scheme in integration with an existing methodology [45] results in a structure, where moving object region labels can be provided in video sequences with good accuracy.

5.3 Limitations

Any framework has its limitations, and our MS benchmark datasets proposal and the semi-automatic motion annotation tools for are no exception. The nature of acquisition of the benchmark video sequences and the annotation tools contain some inherent limitations.

5.3.1 Benchmark limitations

Our benchmark is comprehensively designed to cater for the constraints present in the state-of-the-art datasets. A few limitations of this proposal are;

- Non-rigid motion. The number of non-rigid motions captured in our collection of benchmarks is limited. More specifically, in the non-rigid motion class, only walking motion of people has been acquired. Other motion classes such as animal motion, bird motion or water (liquid) motion, are not present in the datasets and should be included in the future.

- Noisy trajectories. The noisy trajectories in our trajectory-based long sequences dataset are labeled based on the motion they are more representative of. A MS algorithm result might be considered wrong if it classifies the trajectory to be the motion it was less representative of, though it is not completely wrong. A dual label may be a better solution.

5.3.2 Annotation tool limitations

We presented annotation tools to provide ground-truth labels on trajectory- and region-based motion segmentation datasets. A few limitations of our tools are;

- Occlusion limit. The motion-region annotation proposal heavily relies on the optical flow vectors inside the object region. A situation might occur where the occluder overlap with the moving object region covers such an amount of the object that the optical flow vectors' orientation are not dominant in the histogram anymore. This will lead to the deduction of a wrong orientation from the histogram.

- Features in occluder tracking. The use of SURF feature descriptors in occluder mask tracking is dependent upon the the number of SURF features acquired in the occluder mask region. A minimum of three matched features are necessary to formulate a transformation matrix. Increased number of matches would give a better result. If the total number of SURF features and the number of matched features decreases, then the results are effected. This can happen up to a limit where the algorithm might fail, in case less than three matches are found. Even though we did not encounter this issue, but in our experiments it is a point that should be considered.

- Relative change in object size. The overall change of moving object size in each subsequent frame might become a limiting factor. The histogram of optical flow vectors present inside the moving object region gives out the direction of object motion as explained in Section 4.3.1.2. If the object is coming towards the camera, then its size in the FoV is increasing. In this case, in each subsequent frame, the propagated object mask falls inside the actual object region. On the other hand, if the object is going away from the camera, then the object size in each subsequent frame decreases. In this case, the propagated object mask will overlap with the actual object region as well as the background. If the ratio of the background is more than the object, the propagated results might be wrong. This change in size is a limiting factor in object mask propagation.

5.4 Further work

The main directions of further work can be viewed in short-term and long-term perspectives.

The short-term goals include:

► The capability of an algorithm to handle missing data is imperative. All the data captured these days contains missing data in terms of moving objects. Each form of occlusion presents its own set of challenges and give rise to missing data as well. Even if the algorithms are not able to effectively deal with it, they should at least have an inherent structure to process it better. This can be achieved in optical flow based methods where LDOF can be useful. In subspace clustering based methods, the noise margin in intra-cluster affinity should be looked into. Due to increased density of objects in videos, occlusion is a frequently occurring phenomenon that should also be tackled.

- A long term analysis of trajectories may result in a better segmentation of stopping and multiple appearance-disappearance motion. A forward and backward extrapolation based matching of inter-object clusters can help improve the results.

- A few recent proposals of MS algorithms might give better results on our benchmarks. Some promising state-of-the-art algorithms with this potential are described here. In [67], the multiple segment proposal generation on motion boundaries and ranking with a moving objectness notion, establishes affinities for multiple figure-ground segmentation hypothesis. This notion complemented with color, appearance and motion cue used in [39] can yield improved results. Similarly, the minimum cost multicut formulation in [62], optimizes not only for cluster assignments but also for the number of clusters while allowing varying cluster sizes. This technique can complement the results presented in [34]. The capability to combine, potentially imperfect proposals in [61], to improve overall segmentation accuracy and to maintain robustness towards outliers can improve results in the presence of occlusions.

- The background label splitting happens a lot even in small camera motions. A post processing step of spatio-temporal analysis can use affinity measures between labels to merge background labels. Once the background is correctly deciphered from motion, the intra-motion classification can be better performed.

- A natural progression of our region annotation framework is to develop enhanced methods for occluder mask shape tracking. The current method is suitable for rigid shapes with affine transformations. An occluder undergoing non-linear change, like perspective or radial distortion, would be badly tracked by this methodology, as the overall result is sensitive to its shape tracking. A recently proposed shape tracking algorithm [85] might yield better results, as it takes a coarse to fine region-based sobolev descent approach.

- A non-linear scaling adaptation factor can further improve the annotation result on a fine scale. One way of doing it is to perform forward and backward propagation of the object mask, and then to devise a cost function, which penalizes the non-homogeneous region overlap of the mask with the image. Assuming that the homogeneous region is part of the object and non-homogeneous is background, a piece-wise fine scale adjustment of the object mask contour can be done. The objective function in such an approach would be non-linear and computationally extensive, but the results would be better.

Among the long-term work directions we can highlight the following:

- The presented motion segmentation datasets are usable for semi-automatic annotation [74] [88] [89] and semantic segmentation [24] [90] [91]. Our database contains ground-truth on all the frames of all the sequences. The benefit of this complete ground-truth is that this work can also be treated as a dataset for automatic annotation and label propagation algorithms. This research area can take advantage of the diversity in our dataset to test trajectory as well as region label propagation. Even in the case of semantic segmentation, the semantics related to moving objects are already labeled. More semantic labels on static background objects can be easily added through the annotation tools to increase the diversity of our dataset with respect to semantic segmentation problem.

- An enhancement in the object mask propagation approach is needed to deal with non-rigid motion masks. Currently, the motion mask is restricted to being rigid, which is good enough to cater for a lot of real motions but not all. To deal with non-rigid motion masks, the recently proposed scheme of minimal basis subspace based rigid and non-rigid segmentation approach [92] coupled with occlusion-disocclusion segregation [85] can be used in a motion model specific framework to yield acceptable results. A drawback of using image segmentation approaches for moving objects is that based on the number of frames in a video sequence the computational cost multiplies. In comparison, our approach yields quick results depending upon how fast LDOF is being calculated.

- One of the trending frameworks in computer vision are deep learning based methodologies. While discussing the future trends in MS, deep learning framework can not be ignored. Though, the state-of-the-art analysis of deep learning does not include an MS algorithm, there are considerable research proposals presented in related fields. For instance, recent deep learning based methods include human pose estimation [93], human activity recognition [94], gesture detection and localization [95], image segmentation [96], image classification [97], visual tracking [98], video classification [90], etc. A natural progression of these approaches is towards motion segmentation and classification, and semantic video segmentation, which has also an increasing interest within the community. A main limitation in developing deep learning based methods for MS is the requirement of huge training set to train the deep convolutional neural network (CNN) layers in the deep learning framework. The state-of-the-art MS datasets were not able to meet this training-testing data requirement, but with our dataset proposal, a preliminary division of training and testing sets of motions can be done i.e. in region based long sequences: 120 motions in training set and 115 motions in testing set. With this division a deep CNN framework can be designed, trained and tested to propose an initial benchmark on motion segmentation. A proposal like this can pave the way for future advancements in MS based on deep learning. On the same lines for motion annotation, a recent object detection and segmentation approach based on convolutional networks [99] exhibits excellent results. This approach applied integrated with deep networks based object recognition methodologies [100, 101] can yield improved results. These too would work at an exceptionally high computational cost, with a disadvantage of training and testing cycle as necessary for these approaches.

References

[1] D. Song, C. Kim, and S.-K. Park, "A multi-temporal framework for high-level activity analysis: Violent event detection in visual surveillance," *Information Sciences*, vol. 447, pp. 83–103, 2018. 1

[2] H. Yao, A. Cavallaro, T. Bouwmans, and Z. Zhang, "Guest editorial introduction to the special issue on group and crowd behavior analysis for intelligent multicamera video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 405–408, 2017. 1

[3] I. Huerta, M. Pedersoli, J. Gonzàlez, and A. Sanfeliu, "Combining where and what in change detection for unsupervised foreground learning in surveillance," *Pattern Recognition*, vol. 48, no. 3, pp. 709–719, 2015. 1, 2

[4] A. K. S. Kushwaha and R. Srivastava, "A framework of moving object segmentation in maritime surveillance inside a dynamic background," *Journal of Computational Science*, pp. 35–54, 2015. 1, 2

[5] M. N. Ali, M. Abdullah-Al-Wadud, and S.-L. Lee, "Multiple object tracking with partial occlusion handling using salient feature points," *Information Sciences*, vol. 278, pp. 448–465, 2014. 1

[6] L. Wei, X. Wang, J. Yin, and A. Wu, "Self-regularized fixed-rank representation for subspace segmentation," *Information Sciences*, vol. 412, pp. 194–209, 2017. 1

[7] A. K. KC, L. Jacques, and C. De Vleeschouwer, "Discriminative and efficient label propagation on complementary graphs for multi-object tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 61–74, 2017. 1

[8] B.-J. Chen and G. Medioni, "Exploring local context for multi-target tracking in wide area aerial surveillance," *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 787–796, 2017. 1

[9] C. Rubino, M. Crocco, V. Murino, and A. Del Bue, "Semantic multi-body motion segmentation," *IEEE Winter Conference on Applications of Computer Vision*, pp. 1145–1152, 2015. 1, 2

[10] W. Liu, R. Lau, and D. Manocha, "Robust individual and holistic features for crowd scene classification," *Pattern Recognition*, vol. 58, pp. 110–120, 2016. 1

[11] Y. Li, X. Wang, W. Liu, and B. Feng, "Deep attention network for joint hand gesture localization and recognition using static rgb-d images," *Information Sciences*, vol. 441, pp. 66–78, 2018. 1

[12] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, 2016. 1

[13] I. Mademlis, A. Tefas, and I. Pitas, "A salient dictionary learning framework for activity video summarization via key-frame extraction," *Information Sciences*, vol. 432, pp. 319–331, 2018. 1

[14] P. Voigtlaender and B. Leibe, "Online adaptation of convolutional neural networks for video object segmentation," *arXiv preprint arXiv:1706.09364*, 2017. 1

[15] J. Pont-Tuset, S. Caelles, F. Perazzi, A. Montes, K.-K. Maninis, Y. Chen, and L. Van Gool, "The 2018 davis challenge on video object segmentation," *arXiv preprint arXiv:1803.00557*, 2018. 1

[16] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool, "One-shot video object segmentation," *arXiv:1611.05198*, 2017. 1

[17] P. Bideau and E. Learned-Miller, "It's moving! a probabilistic model for causal motion segmentation in moving camera videos," *European Conference on Computer Vision*, pp. 433–449, 2016. 1, 74, 77, 79

[18] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid, "Learning to detect motion boundaries," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2578–2586, 2015. 1, 74, 77, 79

[19] C. Shen, Y. Chen, and X. Guan, "Performance evaluation of implicit smartphones authentication via sensor-behavior analysis," *Information Sciences*, vol. 430, pp. 538–553, 2018. 1

[20] S. Yi, H. Li, and X. Wang, "Understanding pedestrian behaviors from stationary crowd groups," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3488–3496, 2015. 1

[21] D. Yang, J. Guo, Z.-J. Wang, Y. Wang, J. Zhang, L. Hu, J. Yin, and J. Cao, "Fastpm: An approach to pattern matching via distributed stream processing," *Information Sciences*, vol. 453, pp. 263–280, 2018. 1

[22] L. Liu, S. Wang, G. Su, B. Hu, Y. Peng, Q. Xiong, and J. Wen, "A framework of mining semantic-based probabilistic event relations for complex activity recognition," *Information Sciences*, vol. 418, pp. 13–33, 2017. 1

[23] Y. Zhang, H. Lu, L. Zhang, and X. Ruan, “Combining motion and appearance cues for anomaly detection,” *Pattern Recognition*, vol. 51, pp. 443–452, 2016. 1

[24] Y. Zhang, X. Chen, J. Li, C. Wang, and C. Xia, “Semantic object segmentation via detection in weakly labeled video,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3641–3649, 2015. 2, 94

[25] B. Taylor, V. Karasev, and S. Soatto, “Causal video object segmentation from persistence of occlusions,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4268–4276, 2015. 2, 17, 20, 39, 41

[26] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3395–3402, 2015. 2, 17

[27] H. Jung, J. Ju, and J. Kim, “Rigid motion segmentation using randomized voting,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1210–1217, 2014. 2, 17

[28] D. Varas and F. Marques, “Region-based particle filter for video object segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3470–3477, 2014. 2, 17

[29] M. Narayana, A. Hanson, and E. Learned-Miller, “Coherent motion segmentation in moving camera videos using optical flow orientations,” *IEEE International Conference on Computer Vision*, pp. 1577–1584, 2013. 2, 17

[30] J. Yan and M. Pollefeys, “A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate,” *European Conference on Computer Vision*, vol. 3954, pp. 94–106, 2006. 2, 17

[31] E. Elhamifar and R. Vidal, “Sparse subspace clustering: Algorithm, theory, and applications,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, pp. 2765–2781, November 2013. 2, 7, 14, 17, 38, 90

[32] G. Liu, Z. Lin, and Y. Yu, “Robust subspace segmentation by low-rank representation,” *International Conference on Machine Learning*, pp. 663–670, 2010. 2, 17, 41

[33] G. Liu and S. Yan, “Latent low-rank representation for subspace segmentation and feature extraction,” *International Conference on Computer Vision*, pp. 1615–1622, 2011. 2, 7, 17, 38, 41, 90

[34] L. Zappella, X. Lladó, E. Provenzi, and J. Salvi, “Enhanced local subspace affinity for feature-based motion segmentation,” *Pattern Recognition*, vol. 44, no. 2, pp. 454–470, 2011. 3, 7, 19, 38, 41, 90, 93

[35] L. Zappella, E. Provenzi, X. Lladó, and J. Salvi, “Adaptive motion segmentation algorithm based on the principal angles configuration,” *Asian Conference on Computer Vision*, pp. 15–26, 2011. 3, 19, 38

[36] L. Zappella, A. D. Bue, X. Lladó, and J. Salvi, "Joint estimation of segmentation and structure from motion," *Computer Vision and Image Understanding*, vol. 117, no. 2, pp. 113–129, 2013. 3, 19

[37] S. Li, K. Li, and Y. Fu, "Temporal subspace clustering for human motion segmentation," *IEEE International Conference on Computer Vision*, pp. 4453–4461, 2015. 3, 19, 39

[38] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," *Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007. 3, 4, 12, 13, 14, 18, 35, 38, 48, 51

[39] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187–1200, 2014. 3, 4, 7, 12, 14, 18, 20, 34, 35, 38, 39, 41, 48, 52, 57, 59, 72, 90, 93

[40] A. S. Georghiades, P. N. Belhumeur, and D. J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 643–660, 2001. 3, 17

[41] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label mrf optimization," *International Journal of Computer Vision*, vol. 100, no. 2, pp. 190–202, 2012. 3, 16, 18, 35

[42] J. Fauqueur, G. Brostow, and R. Cipolla, "Assisted video object labeling by joint tracking of regions and keypoints," *International Conference on Computer Vision*, pp. 1–7, 2007. 3, 17, 35

[43] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories," *Transcations on Pattern Analysis and Machine Intelligence*, vol. 32, no. 10, pp. 1832–1845, October 2010. 7, 19, 38, 41, 90

[44] V. M. Patel, H. Van Nguyen, and R. Vidal, "Latent space sparse subspace clustering," *International Conference on Computer Vision*, pp. 225–232, 2013. 7, 41, 90

[45] C. Liu, W. T. Freeman, E. H. Adelson, and Y. Weiss, "Human-assisted motion annotation," *Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. 7, 22, 23, 26, 36, 62, 64, 70, 83, 88, 89, 91

[46] T. Brox and J. Malik, "Large displacement optical flow: descriptor matching in variational motion estimation," *Transcations on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 500–513, March 2011. 8, 32, 52, 68

[47] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008. 8, 66

[48] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010. 8, 69

[49] B. Jian and B. C. Vemuri, “Robust point set registration using gaussian mixture models,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1633–1645, 2011. 8, 69

[50] T. Brox and J. Malik, “Object segmentation by long term analysis of point trajectories,” *European Conference on Computer Vision*, pp. 282–295, 2010. 14, 20

[51] W. Chiu and M. Fritz, “Multi-class video co-segmentation with a generative multi-video model,” *Conference on Computer Vision and Pattern Recognition*, pp. 321–328, 2013. 15, 18, 48

[52] F. Li, T. Kim, A. Humayun, and J. M. Tsai, D.and Rehg, “Video segmentation by tracking many figure-ground segments,” *International Conference on Computer Vision*, pp. 2192–2199, 2013. 16, 18, 35, 48

[53] F. Galasso, N. S. Nagaraja, T. Jimenez Cardenas, T. Brox, and B. Schiele, “A unified video segmentation benchmark: Annotation, metrics and analysis,” *International Conference on Computer Vision*, pp. 3527–3534, 2013. 17

[54] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, and J. Malik, “Occlusion boundary detection and figure/ground assignment from optical flow,” *Conference on Computer Vision and Pattern Recognition*, pp. 2233–2240, 2011. 17

[55] R. Vidal, Y. Ma, and S. Sastry, “Generalized principal component analysis (gpca),” *Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945–1959, December 2005. 19, 38

[56] N. Thakoor, J. Gao, and V. Devarajan, “Multibody structure-and-motion segmentation by branch-and-bound model selection,” *Transactions on Image Processing*, vol. 19, no. 6, pp. 1393–1402, June 2010. 19, 38

[57] Y. Wang, J. Gong, D. Zhang, C. Gao, J. Tian, and H. Zeng, “Large disparity motion layer extraction via topological clustering,” *Transactions on Image Processing*, vol. 20, no. 1, pp. 43–52, 2011. 19

[58] K. Nordberg and V. Zografos, “Multibody motion segmentation using the geometry of 6 points in 2d images,” *Proceedings of the International Conference on Pattern Recognition*, pp. 1783–1787, 2010. 19

[59] F. Xu, K. Lam, and Q. Dai, “Video-object segmentation and 3d-trajectory estimation for monocular video sequences,” *Image and Vision Computing*, vol. 29, no. 2, pp. 190–205, 2011. 19

[60] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004. 19

[61] F. Perazzi, O. Wang, M. Gross, and A. Sorkine-Hornung, “Fully connected object proposals for video segmentation,” *IEEE International Conference on Computer Vision*, pp. 3227–3234, 2015. 20, 41, 93

[62] M. Keuper, B. Andres, and T. Brox, "Motion trajectory segmentation via minimum cost multicuts," *IEEE International Conference on Computer Vision*, pp. 3271–3279, 2015. 20, 41, 93

[63] L. Wen, D. Du, Z. Lei, S. Li, and M. Yang, "Jots: Joint online tracking and segmentation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2226–2234, 2015. 20, 39

[64] A. Milan, L. Leal-Taixé, K. Schindler, and I. Reid, "Joint tracking and segmentation of multiple targets," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5397–5406, 2015. 20, 39

[65] P. Ochs and T. Brox, "Higher order motion models and spectral clustering," *Conference on Computer Vision and Pattern Recognition*, pp. 614–621, 2012. 20, 33, 34, 39, 40, 42

[66] ——, "Object segmentation in video: a hierarchical variational approach for turning point trajectories into dense regions," *International Conference on Computer Vision*, pp. 1583–1590, 2011. 20

[67] K. Fragkiadaki, P. Arbeláez, P. Felsen, and J. Malik, "Learning to segment moving objects in videos," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4083–4090, 2015. 20, 41, 93

[68] Y. Yang, G. Sundaramoorthi, and S. Soatto, "Self-occlusions and disocclusions in causal video object segmentation," *IEEE International Conference on Computer Vision*, pp. 4408–4416, 2015. 20, 39

[69] D. Doermann and D. Mihalcik, "Viper: Tools and techniques for video performance evaluation applied to scene and document images," *Symposium on Document Image Understanding Technology*, p. 339, 2001. 21, 22

[70] D. Mihalcik and D. Doermann, "The design and implementation of viper," *University of Maryland*, 2003. 21, 22

[71] I. Kavasidis, S. Palazzo, R. Di Salvo, D. Giordano, and C. Spampinato, "A semi-automatic tool for detection and tracking ground truth generation in videos," *1st International Workshop on Visual Interfaces for Ground Truth Collection in Computer Vision Applications*, p. 6, 2012. 21

[72] ——, "An innovative web-based collaborative platform for video annotation," *Multimedia Tools and Applications*, vol. 70, no. 1, pp. 413–432, 2014. 21

[73] J. Yuen, B. Russell, C. Liu, and A. Torralba, "Labelme video: Building a video database with human annotations," *IEEE 12th International Conference on Computer Vision*, pp. 1451–1458, 2009. 22

[74] S. Bianco, G. Ciocca, P. Napoletano, and R. Schettini, "An interactive tool for manual, semi-automatic and automatic video annotation," *Computer Vision and Image Understanding*, vol. 131, pp. 88–99, 2015. 22, 94

[75] V. Badrinarayanan, F. Galasso, and R. Cipolla, "Label propagation in video sequences," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3265–3272, 2010. 23, 33, 35

[76] I. Budvytis, V. Badrinarayanan, and R. Cipolla, "Label propagation in complex video sequences using semi-supervised learning." *British Machine Vision Conference*, vol. 2257, pp. 2258–2259, 2010. 23, 33, 35

[77] A. Chen and J. Corso, "Propagating multi-class pixel labels throughout video frames," *Western New York Image Processing Workshop*, pp. 14–17, 2010. 23

[78] S. Vijayanarasimhan and K. Grauman, "Active frame selection for label propagation in videos," *European Conference on Computer Vision*, pp. 496–509, 2012. 23, 24, 33, 35, 36

[79] V. Karasev, A. Ravichandran, and S. Soatto, "Active frame, location, and detector selection for automated and manual video annotation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2123–2130, 2014. 23, 24

[80] L. Gao, J. Song, F. Nie, Y. Yan, N. Sebe, and H. Tao Shen, "Optimal graph learning with partial tags and multiple features for image and video annotation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4371–4379, 2015. 24

[81] K. In Kim, J. Tompkin, H. Pfister, and C. Theobalt, "Context-guided diffusion for label propagation on graphs," *IEEE International Conference on Computer Vision*, pp. 2776–2784, 2015. 24

[82] Y. Liu, Z. Yuan, B. Chen, J. Xue, and N. Zheng, "Illumination robust color naming via label propagation," *IEEE International Conference on Computer Vision (ICCV)*, pp. 621–629, 2015. 25

[83] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, pp. 1–20, May 2013. 32

[84] N. Sundaram, T. Brox, and K. Keutzer, "Dense point trajectories by gpu-accelerated large displacement optical flow," *European Conference on Computer Vision*, pp. 438–451, 2010. 32

[85] Y. Yang and G. Sundaramoorthi, "Shape tracking with occlusions via coarse-to-fine region-based sobolev descent," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 5, pp. 1053–1066, 2015. 86, 93, 94

[86] E. Elhamifar and R. Vidal, "Sparse subspace clustering," *Conference on Computer Vision and Pattern Recognition*, pp. 2790–2797, 2009. 90

[87] S. Rao, R. Tron, R. Vidal, and Y. Ma, "Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories," *Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008. 90

[88] Y. Wu, M. Pei, M. Yang, J. Yuan, and Y. Jia, "Robust discriminative tracking via landmark-based label propagation," *IEEE Transactions on Image Processing*, vol. 24, no. 5, pp. 1510–1523, 2015. 94

[89] W. Luo, T.-K. Kim, B. Stenger, X. Zhao, and R. Cipolla, "Bi-label propagation for generic multiple object tracking," *Conference on Computer Vision and Pattern Recognition*, pp. 1290–1297, 2014. 94

[90] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," *Conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014. 94, 95

[91] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014. 94

[92] C. M. Lee and L. F. Cheong, "Minimal basis subspace representation: A unified framework for rigid and non-rigid motion segmentation," *International Journal of Computer Vision*, pp. 1–25, 2016. 94

[93] A. Jain, Y. Tompson, J.and LeCun, and C. Bregler, "Modeep: A deep learning framework using motion features for human pose estimation," *Asian Conference on Computer Vision*, pp. 302–315, 2014. 94

[94] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2013. 94

[95] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," *Workshop at the European Conference on Computer Vision*, pp. 474–490, 2014. 94

[96] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *International Conference on Learning Representations*, pp. 234–247, 2015. 94

[97] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012. 94

[98] Y. Xue, H.and Liu, D. Cai, and X. He, "Tracking people in rgbd videos using deep learning and motion clues," *Neurocomputing*, vol. 204, pp. 70–76, 2016. 95

[99] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016. 95

[100] Z. Wang, Z.and Deng and S. Wang, “Sam: A rethinking of prominent convolutional neural network architectures for visual object recognition,” *2016 International Joint Conference on Neural Networks*, pp. 1008–1014, 2016. 95

[101] L. A. Alexandre, “3d object recognition using convolutional neural networks with transfer learning between input channels,” *Intelligent Autonomous Systems 13*, pp. 889–898, 2016. 95