A Review of Motion Segmentation: Approaches and Major Challenges

Jana Mattheus[1]
[1]*Department of. Electrical Electronic and Computer Engineering*
University of Pretoria
South Africa
mattheus.jana@gmail.com

Hans Grobler[1]
[1]*Department of Electrical Electronic and Computer Engineering*
University of Pretoria
South Africa
hans.grobler@up.ac.za

Adnan M. Abu-Mahfouz[1,2]
[2]*Council for Scientific and Industrial Research*
South Africa
a.abumahfouz@ieee.org

***Abstract*—Motion segmentation has applications in, amongst others, robotics, traffic monitoring, sports analysis, inspection, video surveillance, compression, and video indexing. However, the performance of most methods is limited compared to human capabilities. Based on extensive literature the following challenges remain: occlusions, temporary stopping, missing data, and segmenting multiple objects. In this paper, several popular and state-of-the-art methods were reviewed, with the focus on the most important attributes. These methods were classified according to the main approach taken, namely Image Difference, Optical Flow, Wavelet, Statistical, Layers, Manifold Clustering, Template Matching, and Deep Learning. The investigated methods are compared and major research challenges are highlighted. Based on the review, improvements are identified as a basis for future research.**

Keywords—Motion Segmentation, Motion Analysis, Articulated, Non-rigid, Factorization Method, Manifold Clustering, 3D Scene Analysis, Computer Vision

I. INTRODUCTION

Motion segmentation attempts to extract the moving objects from a video sequence [1]. Motion segmentation is used in applications such as robotics, traffic monitoring, sports analysis, inspection, video surveillance, compression, and video indexing. Since motion segmentation is used for a vast number of applications, the problem is ill-posed and is dependent on the application [2], [3]. Another important consideration is that of the applicable priors [4], [5]. If the expected types of motion are known it changes the approach significantly, and/or could be used to constrain the solution space. For example, in traffic surveillance applications, the aim of motion segmentation is to extract each individual vehicle and possibly pedestrians. Therefore, independent motions can be assumed and the solution space is significantly reduced [6]. However, in sports analysis, the aim of motion segmentation can be to extract the body parts of an athlete for analysis of his/her sport technique. Therefore, the motions can be assumed to be articulate and partially dependent. This also results in a reduction of the solution space but is a more challenging problem than solving for independent motions only.

Occlusions occur when parts of the scene are obscured by objects and methods must be able to handle these situations [7], [8]. Some methods are only able to handle partial occlusions, and fail when a moving object is completely occluded [4], [9]–[11]. Occlusions, noise, changes in lighting conditions, and objects leaving the scene are factors that lead to *missing data*. Motion segmentation methods need to prevent missing data from negatively influencing the accuracy of the extracted motion segments [9], [12]. Another challenge encountered when segmenting motion is the ability to handle the *temporary stopping* of dynamic objects [11], [13]. This is a challenging problem and many methods fail to segment objects that remain stationary for a number of frames. To produce connected and consistent motion segments, *spatial continuity* is exploited [2], [5], [8], [12], [14], [15]. Each pixel is not considered on its own, but its immediate neighborhood is taken into account. A scene can contain *multiple objects* and methods must be able to segment these motions simultaneously [1], [16]. Some methods are limited in the number of motions that can be segmented accurately [10], [11].

Motion can be described by two attributes, namely the dependency and the type. The *dependency* refers to the ability of a method to segment independent, dependent, or partially dependent motions, and describes the relationship between the motions encountered within a scene. The *type* refers to the kind of motion that a method can handle: rigid, non-rigid, articulated, and degenerate [9]. A segment exhibits *rigid* motion when the relative distance between its points as well as their relative position stays the same. A weighted set of rigid base shapes are used to estimate *non-rigid* motion regions [17]. *Articulated* motion can be defined as two or more parts that exhibit dependent motion and are connected by a link such as a joint or an axis [4], [5], [18]. *Degenerate* motion is caused by degenerate objects and has a subspace dimension of lower than the theoretical maximum.

The literature on motion segmentation is extensive and several different approaches can be used to solve different application-specific motion segmentation problems. However, despite the research efforts, the performance of most methods remains limited compared to human capability [19]. Previous review papers gave an overview of current state-of-the-art motion segmentation methods and provided an insight into which direction research was heading [13], [20]. In this paper, popular and current state-of-the-art methods were investigated and classified according to the main approach used. Then, the major research challenges associated with these approaches were highlighted. Methods that aim to provide a solution for similar definitions of motion segmentation were compared and discussed. Finally, enhancements are identified as a basis for developing a novel motion segmentation method.

II. MOTION SEGMENTATION APPROACHES

A number of popular and state-of-the-art motion segmentation methods were investigated and are discussed next. Since the motion segmentation literature is so large,

only sixteen methods were investigated, giving focus to different approaches that have been used in different years. These methods were roughly categorized according to the main principle followed. However, some of these methods fall into more than one category. The categories are as follows: Image Difference, Optical Flow, Wavelet, Statistical, Layers, Manifold Clustering, Template Matching, and Deep Learning.

A. Image Difference

Image difference is a simple approach to detecting changes in a video sequence. Pixel-wise thresholding is applied to the intensity difference between consecutive frames [1], [3], [21]. These methods detect independent motions and can handle occlusions, non-rigid, and articulated motion but are sensitive to noise and changes in lighting conditions, temporary stopping objects, and moving cameras [11]. The method in [3], uses the pixel variance and covariance to model the background. A statistical model of the background is learnt by observing a number of frames and computing the variance and covariance. The variance estimates the absolute variation of the pixel intensity while the covariance estimates the variations of the pixel intensity values relative to that of other pixels. The hue and saturation of the pixel HSV intensity value are used to determine the locations of object shadows that are removed. Motion segmentation is achieved by performing average frame differencing. Morphological operations are used to refine the computed segments. The use of average frame differencing allows the execution time to be near real-time. The background can be redetermined which makes the method robust to dynamic backgrounds. It is also robust to changes in lighting conditions and can handle occlusions.

B. Optical Flow

Methods utilizing optical flow are some of the earliest techniques of analyzing motion in a video sequence. Optical flow on its own cannot be used to segment motion since it fails when encountering occlusions or temporary stopping [2], [15], [22]. It is also sensitive to noise and changes in lighting conditions. Therefore, additional procedures are needed to recover accurate motion boundaries for these two cases [11]. In [23], a Convolutional Neural Network (CNN) framework relies on optical flow to achieve motion segmentation. A Motion Pattern Network (MP-Net) uses the optical flow to separate the camera motion from that of the independently moving object. MP-Net has an encoder-decoder Fully Convolutional Network (FCN) architecture and assigns two labels to each pixel. Noise reduction is employed using object cues and Conditional Random Field (CRF).

C. Wavelet

Wavelet-based motion segmentation methods utilize the properties of wavelets to analyze frequency components of video frames [12], [14], [24], [25]. These methods provide good solutions but can only segment simple motions, such as translations and rotations, and stationary cameras [11]. The methods in [12], [24] use the Discrete Cosine Transform (DCT) to segment motion. The method in [14] relies on Daubechies complex wavelet transform to detect double changes. The wavelet decomposition of three successive frames is obtained by applying Daubechies complex wavelet transform. Then, double change detection is used in the complex wave domain and noise is removed. The edges in the three frames are detected with Canny edge detection, and an edge map is constructed to determine the inter-frame edges and moving edges. Using the moving edges, the moving objects are determined, and a binary closing morphological operation is applied to deal with disconnected edges. Double change detection allows new objects which enter the scene to be detected. The use of Daubechies complex wavelet transform results in better edge detection than methods that rely on real-valued wavelet transforms and reduces the shift sensitivity. The method was found to have higher accuracy and lower misclassification error than other wavelet methods. However, the method fails for cases where the background is nonstationary.

D. Statistical

The motion segmentation problem can be considered a classification problem that aims to classify each pixel as a moving part or as part of the background. In most cases, statistical methods use dense representations, therefore every pixel is classified during the segmentation process. Statistical-based techniques can handle multiple motions as well as occlusions and temporary stopping of motions. The performance of these methods is highly dependent on the motion model and fails if the model is unable to reflect real-world scenarios. In some instances, prior knowledge, such as the number of motions, is required. However, many methods have derived a process to estimate and refine any parameters needed *a priori* [2], [26], [27]. The three commonly used statistical frameworks for motion segmentation are Maximum A Posteriori Probability (MAP), Particle Filter (PF), and Expectation Maximization (EM).

MAP: MAP is based on Bayes rule and classifies pixels to segments such that the posterior probability is maximized [2], [28]. The MAP method presented in [2] combines color and motion segmentation into one framework. For each frame, the segments are determined using the segments of the previous frame. Temporal consistency is imposed using the spatial location and span of every segment as features. This increases the likelihood of obtaining the solution with the smallest change in segment location. The spatial location is combined with the color and Lucas-Kanade optical flow of each pixel to create the feature vector. Weights are selected such that the errors at motion boundaries, due to optical flow and temporal inconsistency, are minimized. The MAP process requires an initial segmentation that is obtained by segmenting the first frame using an EM scheme. The EM scheme computes a Gaussian mixture model that fits the data, namely the color and optical flow, of the first frame. The resulting segments are refined. The method produces accurate segment boundaries and good temporal consistency. However, temporal consistency problems occur when repeated temporary stopping is encountered.

1) Particle Filter: Particle filters track the evolution of a variable over time to compute a sample-based representation of the probability density function and is often adapted to solve the motion segmentation problem [26], [29]. The particle filter-based method in [26] tracks deforming objects. Geometric active contours are used to provide a framework that is parameterization independent while allowing for topology changes. A prior system model and an observational model are included alongside the particle filter. The particle filter estimates the conditional probability distribution of the group action (affine transformation of the

trajectories) and the geometric active contour at a given time instance, conditioned on all the observations up to the given time instance. The standard particle filter is adapted to include Importance Sampling (IS) density, which can be viewed as an approximation to the optimal IS density when the optimal density is multimodal. When the IS variance is very small, indicating a small local deformation of an object, it is replaced by deterministic assignment. Therefore, sampling occurs in the 6D space of affine deformations while local deformations are estimated using the mode of its posterior. The result is that the PF can be executed near real-time. The method also requires significantly fewer particles than other state-of-the-art particle filter methods, and small partial occlusions can be handled.

2) Expectation Maximization: The EM algorithm is used to calculate the Maximum Likelihood (ML) estimate when hidden data are present, or there is missing data [15], [16], [27]. During the E-step of the EM algorithm, conditional expectation is employed to estimate the missing data. Then, the likelihood function is maximized during the M-step. In [15], EM is used to optimize an adapted version of the α-expansion function, which includes label costs. The method in [27] extends the classic dynamic texture model to include a mixture of dynamic textures and is used to segment challenging motions such as smoke, fire, and water from a set of videos. The mixture of dynamic textures model consists of a set of Linear Dynamic Systems (LDS's) that models a set of video sequences. The classic dynamic texture model is extended by adding a hidden variable with the same number of states as the number of textures. This hidden variable describes the spatio-temporal volume that each texture occupies in the video. The volume of the video is modeled as an LDS, conditioned to the hidden variable. During the training stage, the parameters of the mixture of dynamic texture model are computed using an EM algorithm. Segmentation is achieved by clustering the spatial-temporal patches. Simple motions, as well as challenging object motions such as fire, smoke, and traffic, can be modeled accurately. This method suffers from the same drawbacks as LDS and can cause suboptimal solutions to be obtained.

E. Layers

Layered approaches divide video frames into layers according to the number of uniform motions [7], [8]. Each layer has an associated depth and uniform motion parameters as well as parameters that define the motion visibility (i.e. indicates any occlusions). Layer-based methods are highly complex with long execution times but are an effective solution to the occlusion problem [11]. In [8], a layered-based method is presented. A binary mask is computed for each video frame to represent each motion segment. Loopy belief propagation is applied to these masks to compute an initial estimate. Then, αβ-swap and α-expansion are applied iteratively to refine the initial segment and obtain the final motion segments and layering. The αβ-swap and α-expansion algorithms are guaranteed to find a good local minimum of the cost minimization function. After the motion segments have been computed, the texture of each segment is determined by the color values of the pixels belonging to the specific segment. Since the method does not rely on a video frame for an initial estimate, a model can be created for any number of different objects. The method can handle complete occlusions and camera motion.

F. Manifold Clustering

Manifold clustering approaches project the data to a lower-dimensional subspace that preserves some of the properties, such as geodesic distance, of the high-dimensional space. These methods can handle a variety of motion types and temporary stopping [11]. Different manifold clustering approaches exist. These approaches can be classified as Factorization, Subspaces, Sparse Subspace Clustering (SSC), Agglomerative Lossy Compression (ALC), Random Sample Consensus (RANSAC), Generalized Principal Component Analysis (GPCA) and Local Subspace Affinity (LSA).

1) Factorization: Factorization techniques can be applied to segment points residing in different subspaces [4], [9]. A factorization-based approach to segment articulation motion with possible non-rigid parts is presented in [4]. The motion of the articulated and non-rigid parts are modeled as a set of intersecting motion subspaces. The linked parts have 1D and 2D subspace intersection for a joint and axis, respectively. Local sampling and spectral clustering are used to segment the subspaces irrespective of their dimensionality and any dependencies between them. Additionally, the kinematic chain is constructed by considering the intersections between each motion subspace pair, and a minimum spanning tree is used for the computation. After motion segmentation, factorization methods are used to compute the shape of each articulate part, both rigid and non-rigid. The method is robust to outliers. Occlusions and changes in lighting cause incomplete trajectories and the method is unable to handle these cases. Another disadvantage is that the method is based on affine projections, but can be used as an initialization for a perspective projection approach.

2) Subspaces: The properties of points residing in different subspaces can be exploited to achieve motion segmentation [5], [30]. In [5], a piecewise approach is used to 3D reconstruct an articulated object from point trajectories. To reconstruct the articulated object in 3D, a constraint is introduced which forces the segments of neighboring segments to overlap. These overlapping regions are used to stitch the 3D models of each segment together. Therefore, the problem can be formulated as a model assignment problem where every model is associated with an articulated part of the object. The labeling process is optimized by combining graph-cut based inference and Structure from Motion (SfM) factorization. The re-projection error, subject to the constraint that neighboring points must belong to the same model, is used as the cost minimization function for the model assignment and 3D reconstruction steps. This allows the algorithm to switch between assigning points to models and fitting rigid models to parts in a hill-climbing approach. After the segmentation step, the assumption of rigid motion of the links is relaxed and non-rigid reconstruction is applied to each reconstructed region as a post-processing step. The number of motions is not required beforehand and spatial continuity is exploited.

At least three point tracks on each articulated part is needed for 3D reconstruction. Additionally, at least one of these point tracks must lie on the intersection with another articulated part to stitch the 3D models together. These two constraints are guaranteed by the inference model, given that each point has a minimum of two neighbors.

3) SSC: SSC is a sparse subspace clustering algorithm that computes a similarity matrix by solving a relaxed version of the ℓ_0-minimization problem, namely the ℓ_1-norm [31]. The similarity matrix is used to construct a graph and k-means spectral clustering is applied to segment the data points. SSC expresses each point as a linear or affine combination of points, other than itself, from the same subspace. However, using the ℓ_1-norm as the minimization problem can cause large coefficients of the similarity matrix to be contaminated with large errors. Therefore, a new norm, namely the $\ell_{q,e}$-norm, is presented in [32]. This norm is non-convex and can better approximate the ℓ_0-norm, thus improving the similarity matrix. However, the non-convex nature of the proposed norm makes it difficult to solve. Therefore, a re-weighted Alternating Direction Method of Multipliers (ADMM) is used to solve the non-convex sub-problems.

4) ALC: The ALC method proposed in [6] employs lossy compression to cluster trajectories which lie on multiple subspaces of varying dimensionalities, therefore, it is suited for segmenting mixtures of motions. Rank minimization and sparse representation are exploited to handle corrupted trajectories and incomplete data before segmentation is executed. For data with high dimensionality, clustering algorithms can produce suboptimal segments if the trajectory points do not sufficiently span the subspace. Therefore, two techniques are used to address this problem and improve performance. If only affine motions present, the data is projected onto a 5D subspace. If mixtures of motions are present, the data is projected to a subspace larger than the minimum dimensionality of the subspaces.

5) RANSAC: RANSAC can be employed to extract information from the trajectories which can be used during the segmentation process [18], [33]. RANSAC with priors is an articulated motion segmentation method that uses RANSAC as the base for the segmentation algorithm [18]. First, an affinity matrix is constructed. Priors on the likelihood that each trajectory pair is from the same motion, are estimated and used to construct a sample set. RANSAC is used to segment the trajectories into models with similar motion and shape. The sample set can be increased without significantly increasing the computation time since the priors increase the chances that the points belonging to the same model are contained in the sample set. The use of priors, derived from spectral affinities between each trajectory pair, makes the method efficient. The method can segment independent motions by viewing it as a special case and treating it uniformly. The method cannot handle degenerate shape and motion but can be extended for these cases by adding a model selection method. The method was not tested on complicated articulated motion such as human motion or complex scenes with a variety of motions.

6) GPCA: GPCA is a statistical method that considers a group of subspaces to be an algebraic set [10], [34]. Algebraic geometry is used to determine the algebraic set and segment it into subspaces. GPCA can be used to model underlying manifolds and is often used as the bases for motion segmentation. A GPCA-based method is presented in [10] to segment rigid-body motion in multiple affine views using point correspondences. First, the point trajectories are projected onto a 5D subspace using SVD for complete data, PowerFactorization for missing data, or RANSAC for data containing outliers. Then, Spectral GPCA is used to fit a set of subspaces to the projected trajectories. This process starts by fitting a homogeneous polynomial, which represents every motion subspace, to the projected trajectories. The derivatives of this polynomial are used to obtain a basis for every motion subspace. A similarity matrix is built from the subspace angles, and spectral clustering is applied to cluster the projected trajectories. The method can handle incomplete data. Multiple object motions can be segmented, but the performance deteriorates drastically when a scene contains three or more motions. This is since GPCA uses linear least squares to fit multiple non-linearly related coefficients. Noise and outliers cause the estimated coefficients to be inaccurate.

7) LSA: LSA-based methods estimate the subspaces generated by every trajectory. The distance between each subspace pair is computed and used to construct an affinity matrix [11], [19]. This matrix is clustered to obtain the final segmentation. The disadvantage of LSA is that it cannot handle missing data [11] and is heavily dependent on the rank of the trajectory model [19]. LSA estimates the number of motions using Normalized Cuts, but it is not a reliable approach. Therefore, many LSA implementations that use Normalized Cuts assume that the final number of motions is known beforehand. In [11], Enhanced LSA (ELSA), which addresses the flaws of classic LSA for motion segmentation, is presented. Unlike LSA, ELSA tunes the most sensitive parameters automatically. This is achieved by using Enhanced Model Selection (EMS) for model selection since it automatically adjusts to different numbers of motion and noise conditions. The number of motions is estimated dynamically by applying a process based on Linear Discriminant Analysis (LDA) to compute a threshold for the eigenvalue spectrum of the Symmetric Normalized Laplacian matrix. The rest of the parameters were fixed manually and the results proved to be accurate without tuning these parameters. ELSA outperforms ALC on noisy datasets and has reduced computation time.

G. Template Matching

Template matching approaches identify moving parts by obtaining regions of a frame that match a template [35]. Occlusion handling, non-rigid transformations, changes in lighting conditions, and background noise affect the performance of template matching-based methods and additional procedures are required to handle them. These challenges are addressed in [17] which computes the articulated parts of an object from a set of meshes that correspond to different states of the object. The object is

assumed to be mostly rigid. The meshes are registered using Correlated Correspondence, an unsupervised non-rigid registration method. Then, a graphical model that describes the part composition problem is defined, as well as hidden variables that indicate the part to which each point in the mesh belongs to. Due to the rigidity assumption, all points on a part have the same rigid transformation for each registered mesh. Soft spatial continuity constraints ensure reasonable part composition, therefore nearby points are preferred to be assigned to the same part which is represented as undirected edges in the graphical model. EM is performed on the graphical model by iteratively decomposing the object into parts and determining the location of these parts. The optimal number of parts is also determined. Even though the underlying graphical model is densely connected, a global optimization step can be performed to achieve the part decomposition. This allows many parts to be retrieved while avoiding local maxima. Lastly, the articulated joints are estimated by applying articulation constraints. The method can accurately determine the articulated parts of an object from only a few poses. Non-rigid motion does not significantly affect the performance. The registration and part decomposition can be done in a single step. However, applying these separately allows for global robust inference strategies to be applied during both steps and avoids sub-optimal solutions being obtained.

H. Deep Learning

Deep learning architectures, such as neural networks, can also be employed to extract moving parts from a video [36]–[38]. In many instances, these methods require prior knowledge in the form of a training set that is used to set parameter values during the training stage. In [36], a deep learning method is given that extracts the articulated parts of an object from a set of 3D structures corresponding to different states of the object. The architecture consists of three trained neural networks called Correspondence Proposal, Flow, and Segmentation modules. The Correspondence Proposal module is used to compute the shape correspondences by mapping shape pair geometries to probabilistic point-wise correspondences. This allows differences in geometry and articulation between input shapes to be handled and increases the robustness against noise and missing data. The Flow module is based on the PairNet network and extracts relationships between point sets to translate the correspondences to a deformation flow field. Then, the Segmentation module aggregates the deformation flows into piecewise rigid motions to find the articulated parts, and is based on RecurrentPartExtractionNetwork. To achieve optimal performance, an Iterative Closest Point (ICP) approach is used to alternate between the three modules and is terminated once the magnitude of the deformation flow is minimized. The method generalizes well to previously unseen objects. Noise removal is implemented to make the method robust, and the method can handle partial point clouds.

III. Discussion

Table I provides a summary of the investigated methods with respect to the most important factors. The methods are categorized to Image Difference, Optical Flow, Wavelet, Statistical, Layers, Manifold Clustering, Template Matching, and Deep Learning. Cells containing "N/A" indicates that the factor does not apply to the method.

Image difference-based methods are simple and can handle multiple objects, occlusions, and non-rigid and articulate motions. These methods are unable to extract the non-rigid and articulated parts since only independent motions are extracted. The method in [3] can effectively handle temporary stopping and missing data since the background is modeled using statistical methods. Unlike other image difference methods, it is not susceptible to changes in lighting conditions. Optical flow methods are another simple approach to motion segmentation, but optical flow by itself cannot be used to extract motion segments. The optical flow-based method in [23] overcomes these issues with a CNN and is to able handle occlusions and temporary stopping. It is also the only investigated method that is not affected by changes due to lighting conditions. The use of multiple cameras as well as camera motion is supported. The wavelet approach in [14] can extract the shape of rigid and articulated motion and handle new objects entering the scene. Wavelet-based approaches can use multi-resolution analysis to compute depth planes that can be used to solve the occlusion problem.

Statistical methods use dense motion representations and are robust when the model reflects realistic situations. These methods usually require some type of prior knowledge, but the methods in [2], [26] and [27] incorporate methods to estimate these parameters, therefore eliminating the need for prior knowledge. The two methods in [2] and [26] can handle multiple objects, occlusions, and temporary stopping. The method in [27] is focused on segmenting difficult object motion such as fire and smoke, and therefore the characteristics of the method differ from the other two methods. However, a training set is required to set the model parameters beforehand.

The main focus of layered approaches is to solve the occlusion problem. However, these methods are complex with long execution times. Manifold clustering-based methods use key point trajectories to segment motion which allows partial occlusions to be handled. Further, these approaches naturally connect to SfM which allows the 3D structure of the object and camera motion to be obtained. These methods can segment independent and dependent motion as well as extract the articulated parts of an object. The methods in [6], [10], [11], [18], [31] can segment multiple objects. Partial occlusion handling is included in [6], [10], [11], [31]. The methods in [4], [6], [11], [18], [31] all require prior knowledge.

Template matching methods obtain the motion segments by comparing the input of the scene to a template and computing the similar parts. These methods are dependent on the quality of the template, e.g. if a segment is occluded in the template, it will not be recovered. However, if all segments are visible in the template, these methods can handle complete occlusions, since the segments will be redetected once they come into full view again. The template matching method in [17] receives 3D models of the object as input and can extract independent and dependent motions. Deep learning approaches use machine learning techniques such, as neural networks or convolutional neural networks, to solve the segmentation problem. These methods can be trained to extract motion segments with high precision and speed, but often require a training step. The training stage highly influences the performance of these methods. The deep learning method in [36] can segment independent and

TABLE I. SUMMARY OF MOST IMPORTANT ATTRIBUTES OF INVESTIGATED METHODS

Approaches		References	Motion Representation (D Dense, F Feature)	Input Data	Prior Knowledge (N Cluster number, T Training, S Subspace dimension)	Moving Camera	Multiple Cameras	Occlusions	Missing Data	Temporary Stopping	Spatial Continuity	Multiple Objects	Objects Enter (E) or Leave (L)	Dependency (I Independent, D Dependent, P Partial)	Type (R Rigid, N Non-rigid, A Articulated, D Degenerate)
Image Difference		[3]	D	Video		N	N	Y	Y	Y	Y	Y	EL	I	RA
Optical Flow		[23]	D	Video	T	Y	N	Y	Y	Y	Y	Y	EL	I	RNA
Wavelet		[14]	D	Video		N	N	Y	Y	Y	Y	Y	E	I	RA
Statistical	MAP	[2]	D	Video		Y	N	Y	Y	N	Y	Y	EL	I	RA
	PF	[26]	D	Video		Y	N	Y	Y	N	Y	Y	No	I	RNA
	EM	[27]	D	Video	T	N	N	N	N	Y	Y	Y	EL	I	RNAD
Layers		[8]	D	Video		Y	N	Y	Y	Y	Y	Y	EL	IDP	RNA
Manifold Clustering	Factorization	[4]	F	Trajectories	NS	Y	N	N	N	Y	Y	Y	EL	IDP	RNAD
	Subspaces	[5]	F	Trajectories		Y	N	N	Y	Y	Y	Y	No	IDP	RA
	SSC	[32]	F	Trajectories		Y	N	Y	Y	Y	Y	Y	EL	IDP	RNA
	ALC	[6]	F	Trajectories	N	Y	N	Y	Y	Y	Y	Y	EL	IDP	RNAD
	RANSAC	[18]	F	Trajectories	NS	Y	N	N	N	N	Y	N	N/A	IDP	RNA
	GPCA	[10]	F	Trajectories		Y	Y	Y	Y	Y	Y	Y	EL	IDP	RAD
	LSA	[11]	F	Trajectories	N	Y	N	Y	Y	Y	Y	Y	EL	IDP	RA
Template Matching		[17]	D	3D models		N/A	N/A	Y	Y	N/A	Y	N	N/A	IDP	RNA
Deep Learning		[36]	D	3D point clouds	T	N/A	N/A	Y	Y	N/A	Y	N	N/A	IDP	RA

dependent motions and operates on 3D point clouds of the object under observation. Therefore, an additional process is required to extract 3D point clouds from the raw video before the method can be applied.

From table I and the previous section, it is evident that each approach does not solve a single problem. This is since the motion segmentation problem is ill-defined and the definition is dependent on the application. Therefore, different approaches can be used to solve a single motion segmentation problem.

One aim of motion segmentation can be to extract all independently moving objects in a scene. The methods ideal for this application were presented in the following papers: the image difference-based method in [3], optical flow approach in [23], wavelet approach of [14], and the three statistical methods in [2], [26] and [27]. From table I, these methods can accurately extract the shape of different motion types such as rigid, non-rigid, articulated, and degenerate. The optical flow-based method in [23] and the statistical method in [27] require prior knowledge in the form of a training set. The method in [26] is the only one of these methods which cannot handle objects that enter and leave the scene.

Another focus of motion segmentation is to extract the articulated parts of an articulated object. The motion of each part is dependent on the motion of the object as a whole, and therefore the motions are not independent. As seen in table I, different approaches can be used to achieve this. The following methods can segment complex articulated objects with more than three articulated parts: the manifold clustering approaches in [4]–[6], [18], [31], [32] the template matching approach in [17], and the deep learning method in [36]. Methods such as the manifold clustering approaches in [10], [11] can extract the articulated parts but are limited in the number of parts that can be detected. In addition to extracting the articulated parts, the kinematic chain that describes the motion of the parts relative to each other can also be extracted as is done in [4], [11]. The layered approach in [8] can segment highly articulated objects as well as multiple independently moving objects. It is also the only investigated method that can handle complete occlusions.

Other methods attempt to provide a more generic solution that segments different types of motion such as rigid, non-rigid, and articulated motion. More generic methods include the manifold clustering approaches in [4], [6], [10], [11], [31], [32]. The methods in [4], [5], [18] are unable to handle occlusions effectively. From table I, the methods in [4], [6] can segment more types of motion than the rest of the manifold clustering methods. The GPCA method in [10] can handle temporary stopping, but cannot segment non-rigid motion. The LSA method in [11] was reported to outperform

the GPCA-based method in [10], however, it is unable to segment degenerate motion.

Another interesting motion segmentation problem is the extraction of transparent motion and motion of textures. The approaches in [25], [27] can extract the motion of fire and smoke by using mixtures of dynamic textures. These are the only investigated methods able to segment such motion.

IV. Conclusion

It is evident that there are still many unresolved problems in motion segmentation research. One of these gaps is an effective method to handle occlusions. Most of the investigated methods can only handle small, partial occlusions and can be improved by providing additional procedures to handle large, or complete occlusions. Even though the layered approach in [8] attempts to solve the occlusion problem, it is complex with long execution time, and it is unable to segment non-rigid motion. Only the wavelet and statistical-based method in [25] and [27], respectively, can segment transparent motion such as smoke, therefore, similar procedures can be developed to include this functionality to improve any of the other investigated methods. Some of the investigated articulated motion segmentation methods, such as [10], [11], are unable to segment complex articulated objects. None of the investigated methods provide a generic motion segmentation solution that can segment a mixture of different motion types as well as any number of motions.

For the development of a new motion segmentation method, it is proposed to focus on a generic method that can segment both independent and dependent motion as well as rigid, non-rigid, articulated, and degenerate motion. Since manifold clustering is based on a strong mathematical foundation, motion segments and object structure can be extracted easily, and therefore, it is a good initial point for developing a novel method. From table I, it can be seen that the manifold clustering-based methods are unable to handle large and complete occlusions. Many of the manifold clustering-based methods rely on prior knowledge such as the number of clusters or the subspace dimension which limits the number of motions that can be segmented. Automatic estimation of these parameters can be included to eliminate the need for prior knowledge. Procedures to handle incomplete point trajectories can be developed. One approach is to segment the trajectories up until the time instance when the occlusion occurs. Another approach can be to include depth information such as is done in layered approaches. Alternatively, points that undergo occlusions can be detected and segmented separately from points with complete trajectories to detect motions that undergo complete occlusions.

Acknowledgements

This research was supported by the Council for Scientific and Industrial Research, Pretoria, South Africa, through the Smart Networks collaboration initiative and IoT-Factory Program (Funded by the Department of Science and Innovation (DSI), South Africa).

References

[1] Y. Zhang, B. Luo, and L. Zhang, "Permutation Preference Based Alternate Sampling and Clustering for Motion Segmentation," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 432 – 436, 2018.

[2] S. Khan and M. Shah, "Object Based Segmentation of Video Using Color, Motion and Spatial Information," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, Oct. 2001, pp. 746 – 751.

[3] A. Kushwaha, C. Sharma, M. Khare, O. Prakash, and A. Khare, "Adaptive Real-time Motion Segmentation Technique Based on Statistical Background Model," *The Imaging Science Journal*, vol. 62, pp. 285 – 302, Jun. 2014.

[4] J. Yan and M. Pollefeys, "A Factorization-Based Approach for Articulated Nonrigid Shape, Motion and Kinematic Chain Recovery From Video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 865 – 877, May 2008.

[5] J. Fayad, C. Russell, and L. Agapito, "Automated Articulated Structure and 3D Shape Recovery from Point Correspondences," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 431 – 438, Nov. 2011.

[6] S. Rao, R. Tron, R. Vidal, and L. Yu, "Motion Segmentation in the Presence of Outlying, Incomplete, or Corrupted Trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1832 – 1845, Oct. 2010.

[7]] L. Ge, C. Zhang, Z. Chen, and M. Li, "Optical Flow Estimation from Layered Nearest Neighbor Flow Fields," in *2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, 2018, pp. 1 – 6.

[8] M. Pawan Kumar, P. H. S. Torr, and A. Zisserman, "Learning Layered Motion Segmentations of Video," *International Journal of Computer Vision*, vol. 76, no. 3, pp. 301 – 319, Mar. 2008.

[9] M. Paladini, A. Del Bue, M. Stosic, M. Dodig, J. Xavier, and L. Agapito, "Factorization for Non-rigid and Articulated Structure Using Metric Projections," *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 2898 – 2905, Jun. 2009.

[10] R. Vidal, R. Tron, and R. Hartley, "Multiframe Motion Segmentation with Missing Data Using PowerFactorization and GPCA," *International Journal of Computer Vision*, vol. 79, pp. 85 – 105, Aug. 2008.

[11] L. Zappella, X. Llado, E. Provenzi, and J. Salvi, "Enhanced Local Subspace Affinity for Feature-based Motion Segmentation," *Pattern Recognition*, vol. 44, pp. 454 – 470, Feb. 2011.

[12] F. Shi, Z. Zhou, J. Xiao, and W. Wu, "Robust Trajectory Clustering for Motion Segmentation," in *2013 IEEE International Conference on Computer Vision*, Dec. 2013, pp. 3088 – 3095.

[13] L. Zappella, X. Llado, and J. Salvi, "Motion Segmentation: a Review," in *Frontiers in Artificial Intelligence and Applications*, vol. 184, Jan. 2008, pp. 398 – 407.

[14] M. Khare, R. Srivastava, and A. Khare, "Moving Object Segmentation in Daubechies Complex Wavelet Domain," *Signal Image and Video Processing*, vol. 9, pp. 635 – 650, Mar. 2015.

[15] A. Delong, A. Osokin, H. Isack, and Y. Boykov, "Fast Approximate Energy Minimization with Label Costs," *International Journal of Computer Vision*, vol. 96, pp. 2173 – 2180, Jun. 2010.

[16] J. Stückler and S. Behnke, "Efficient Dense 3D Rigid-Body Motion Segmentation in RGB-D Video," in *BMVC 2013 - Electronic Proceedings of the British Machine Vision Conference 2013*, Jan. 2013, pp. 233 – 245.

[17] D. Anguelov, D. Koller, H. C. Pang, P. Srinivasan, and S. Thrun, "Recovering Articulated Object Models from 3D Range Data," in *UAI '04: Proceedings of the 20th conference on Uncertainty in artificial intelligence*, Jul. 2012, pp. 18 – 26.

[18] J. Yan and M. Pollefeys, "Articulated Motion Segmentation Using RANSAC with Priors," *Proceedings of the 2005/2006 International Conference on Dynamical vision*, vol. 4358, pp. 75 – 85, May 2006.

[19] L. Zappella, X. Llado, and J. Salvi, "Enhanced Model Selection for Motion Segmentation," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 4053 – 4056.

[20] M. Liu, Y. Yan, R. Chen, and H. Wang, "The State-of-the-art Research Progress on Motion Segmentation," in *ACM International Conference Proceeding Series*, Jul. 2014, pp. 345 – 349.

[21] A. Colombari, A. Fusiello, and V. Murino, "Segmentation and Tracking of Multiple Video Objects," *Pattern Recognition*, vol. 40, pp. 1307 – 1317, Apr. 2007.

[22] L. Xu, J. Chen, and J. Jia, "A Segmentation Based Variational Model for Accurate Optical Flow Estimation," in *ECCV 2008: 10th European Conference on Computer Vision*, Oct. 2008, pp. 671 – 684.

[23] P. Tokmakov, K. Alahari, and C. Schmid, "Learning Motion Patterns in Videos," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 531 – 539.

[24] Y. Zhang and Z. He, "Video Object Segmentation of Dynamic Scenes with Large Displacements," *IEICE Transactions on Information and Systems*, vol. E98.D, no. 9, pp. 1719 – 1723, 2015.

[25] M. Cai, X. Lu, X. Wu, and Y. Feng, "Intelligent Video Analysis-based Forest Fires Smoke Detection Algorithms," in *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, Aug. 2016, pp. 1504 – 1508.

[26] Y. Rathi, N. Vaswani, A. Tannenbaum, and A. Yezzi, "Tracking Deforming Objects Using Particle Filtering for Geometric Active Contours," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 8, pp. 1470 – 1475, 2007.

[27] A. B. Chan and N. Vasconcelos, "Modeling, Clustering, and Segmenting Video with Mixtures of Dynamic Textures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 909 – 926, 2008.

[28] C. Zheng and H. Yao, "Segmentation for Remote-sensing Imagery using the Object-based Gaussian-Markov Random Field Model with Region Coefficients," *International Journal of Remote Sensing*, vol. 40, pp. 1 – 32, Jan. 2019.

[29] J. H. Hammer, M. Voit, and J. Beyerer, "Motion Segmentation and Appearance Change Detection-based 2D Hand Tracking," in *2016 19th International Conference on Information Fusion (FUSION)*, 2016, pp. 1743 – 1750.

[30] K. Yücer, O. Wang, A. Sorkine-Hornung, and O. Sorkine-Hornung, "Reconstruction of Articulated Objects from a Moving Camera," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, Dec. 2015, pp. 823 – 831.

[31] E. Elhamifar and R. Vidal, "Sparse Subspace Clustering," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 2790 – 2797.

[32] X. Deng, T. Sun, P. Du, and D. Li, "A Nonconvex Implementation of Sparse Subspace Clustering: Algorithm and Convergence Analysis," *IEEE Access*, vol. 8, pp. 54 741 – 54 750, 2020.

[33] R. Serajeh, A. Mousavinia, and F. Safaei, "Motion Segmentation with Hand Held Cameras Using Structure From Motion," in *2017 Iranian Conference on Electrical Engineering (ICEE)*, 2017, pp. 1569 – 1573.

[34] R. Vidal, Yi Ma, and S. Sastry, "Generalized Principal Component Analysis (GPCA)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 12, pp. 1945 – 1959, 2005.

[35] K. Wattanachote and T. K. Shih, "Automatic Dynamic Texture Transformation Based on a New Motion Coherence Metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 10, pp. 1805 – 1820, Oct. 2016.

[36] L. Yi, H. Haibin, D. Liu, E. Kalogerakis, H. Su, and L. Guibas, "Deep Part Induction from Articulated Object Pairs," *ACM Transactions on Graphics*, vol. 37, Sept. 2018.

[37] Y. Wang, P. Wang, Z. Yang, C. Luo, Y. Yang, and W. Xu, "UnOS: Unified Unsupervised Optical-Flow and Stereo-Depth Estimation by Watching Videos," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8063 – 8073.

[38] P. Tzirakis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Time-series Clustering with Jointly Learning Deep Representations, Clusters and Temporal Boundaries," in *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, 2019, pp. 1 – 5.