3

New Trends in Motion Segmentation

Luca Zappella, Xavier Lladó and Joaquim Salvi
University of Girona, Girona, Spain

1. Introduction

Motion segmentation algorithms aim at decomposing a video in moving objects and background. In many computer vision tasks this decomposition 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. A great number of researchers has focused on the segmentation problem and this testifies the relevance of the topic. However, despite the vast literature, the performance of most of the algorithms still falls far behind human perception. In this chapter a review of the main motion segmentation approaches is presented, with the aim of pointing out their strengths and weaknesses and suggesting new research directions. The main features of motion segmentation algorithms are analysed and a classification of the recent and most important techniques is proposed. The conclusions summarise the review and present a vision on the future of motion segmentation algorithms.

2. State of the art

In this section a complete state of the art review on motion segmentation is presented. First the main problems and attributes of motion segmentation algorithms are analysed. Afterwards, a classification of the different techniques is proposed, describing the most significant works in this field. All the papers revised are summarised in table 1, which offers a compact at-a-glance overview with respect to the attributes presented in the next subsection.

2.1 Problems and attributes

In this subsection the common problems and the most important attributes of motion segmentation algorithms are analysed. Attributes describe in a compact way the assumptions made by an algorithms as well as its limitations and strengths.

One of the first choice that has to be taken when developing a motion segmentation algorithm is the *representation* of the motions: there are *feature-based* and *dense-based* approaches. In feature-based methods, the objects are represented by a limited number of salient points. Most of these methods rely on computing a homography corresponding to the motion of a planar object (Kumar et al, 2008). Features represent only part of an object hence the object can be tracked even in case of partial occlusions. In opposition to feature-based methods there are dense-based methods which do not use sparse points but compute

a pixel-wise motion. The result is a more precise segmentation of the objects but the occlusion problem becomes harder to solve (Kumar et al, 2008).
Motion segmentation algorithms usually exploit temporal continuity. However, using only temporal clues a rather big part of the available information is thrown away and this lack of information can easily lead to problems. This is the reason why some techniques exploit also *spatial continuity*. In these cases each pixel is not considered as a single point but the information provided by its neighbourhood (in terms of spatial proximity) is taken into account. For example, one of the problems that are caused by the use of temporal information only, is the ability to deal with *temporary stopping*. In fact, many techniques Fail to segment when the objects stop moving even for a limited amount of time.
Another common problem of motion segmentation is the fact that objects that move are not always visible. The ability to deal with *missing data* is yet one of the most difficult problems. Missing data can be caused by many factors: presence of noise, occlusions, or feature points that are not in scene for the whole length of the sequence. The presence of noise is another cause of failure. Noise can affect the accuracy in the position of the tracked features, or the amount of outliers (erroneously tracked features). Hence, the *Robustness* of the algorithm against noise is an essential factor to take into account. For simplicity, in this chapter the term ˝robustness“ groups together the ability to deal with all the problems caused by noise, as well as the robustness against initialization (when an initial solution is required).
Another important attribute that has to be analysed is the ability to deal with different types of motion. There is a bit of confusion in the literature when it comes to “type of motion” as people tend to use different adjectives to describe the same property of the motion or the same adjective to describe different properties. Hence, it is important to clarify which is the exact meaning that is given to each adjective in this chapter. A motion can be described in terms of: *dependency* and *kind*.
The first classification is between independent and dependent motions. This is an attribute that describe the relationship between a pair of motions and is not a feature of one single motion. Motions are *independent* if the pairwise intersection of the generated subspaces is the zero vector. On the other hand, motions are *dependent* if the pairwise intersection of the subspaces is not empty. In this case the two motions can be seen as “similar”, the dependency can be partial (which means that the subspaces intersect in some points) or complete (which means that one subspace is completely inside the other) (Rao et al, 2008).
The kind of motion is an attribute of the single motion. A motion is *rigid* when the trajectories generated by the points of a rigid object form a linear subspace of dimensions no more than 4 (Tomasi and Kanade, 1992). It is *non-rigid* if the trajectories generated by the points of a non-rigid object can be approximated by a combination of k weighted key basis shapes, and they form a linear subspace of dimension no more than 3k + 1 (Koterba et al, 2005; Llado et al, 2006). It has to be noted that the ability to deal with non-rigid motions is constrained to when the nonrigid structure has also a rigid motion component during its movement. And finally, a motion is *articulated* when is composed by two dependent motions M1 and M2 connected by a link. If the link is a joint, [R1|T1] and [R2|T2] must have T1 = T2 under the same coordinate system. Therefore, M1 and M2 lie in different linear subspaces but have 1-dimensional intersection. If the link is an axis, [R1|T1] and [R2|T2] must have T1 = T2 and exactly one column of R1 and R2 being the same under a proper coordinate system. So M1 and M2 lie in different linear subspaces but have 2-dimensional intersection (Yan and Pollefeys, 2006).

These are all the attributes, related to the description of motion, that will be taken into account in table 1. However, not always the authors clearly state under which conditions the algorithm would work, therefore the table is filled to the best of our knowledge given the information provided in the cited papers. There would be two more Attributes that, for the sake of completeness, are described here but they are not considered in the table as very few authors clearly explain these aspects. The first is called in this article "degeneracy". Many authors use it when they refer to dependent, non-rigid or articulated motions, but it is used here with a different meaning. Degeneracy is another aspect of a single motion. *Non Degenerate Motion* is a motion whose subspace dimension is the maximum (i.e. 4 for rigid motion, 3k + 1 for nonrigid motion, etc.). Whereas, *Degenerate Motion* is a motion whose subspace have a dimension which is lower than its theoretical maximum due to some degeneracies in the trajectories. The second attribute is the assumed camera model, which can be *affine, perspective, para-perspective or projective*
Furthermore, if the aim is to develop a generic algorithm able to deal in many unpredictable situations there are some algorithm features that may be considered as a drawback. For instance, one important aspect is the amount of *prior knowledge* required. In particular: number of moving objects and dimension of the generated subspaces. A second aspect is the fact that some algorithm require a *training* step. Training is not a negative point itself, however a trained algorithm tends to lose generality and it requires extra effort and a relevant amount of data that is not always available.

2.2 Strategies analysis

As motion segmentation has been a hot topic for many years its literature is particularly wide. In order to make the overview easier to read and to create a bit of order, the approaches will be divided into categories which represent the main principle underlying the algorithm. For each category some articles, among the most representative and the newest proposals, are provided. The division is not meant to be tight, in fact some of the algorithms could be placed in more than one category. The groups identified are: Image Difference, Statistical, Optical Flow, Wavelets, Layers, and Manifolds Clustering. As the amount of literature is notable only the main idea of each group of techniques is described while details about each paper are presented in the table 1.

2.2.1 Image difference

Image difference is one of the simplest and most used techniques for detecting changes. It consists in thresholding the intensity difference of two consecutive frames pixel by pixel. The result is a coarse map of the temporal changes. An example of an image sequence and the image difference result is shown in figure 1. Despite its simplicity, this technique cannot be used in its basic version because it is really sensitive to noise. Moreover, when the camera is moving the whole image changes and, if the frame rate is not high enough, the result would not provide any useful information. Works based on image difference usually focus on these two problems. For example, in (Cavallaro et al, 2005) the authors reinforce the motion difference using a probability-based test in order to change the threshold locally. In (Cheng and Chen, 2006) they exploit the wavelet decomposition in order to reduce the noise. Other proposals, like (Li et al, Aug. 2007), try to use temporal and spatial information simultaneously to be able to deal with noise and to solve other typical

Image		(Cavallaro et al, 2005)	F/D	✓	✓	✓		✓	✓	
		(Cheng and Chen, 2006)	D	✓	✓		✓	✓	✓	X
Diff.		(Li et al, Aug. 2007)	D		✓	✓		✓	✓	
		(Colombari et al, 2007)	D	✓	✓	✓	✓	✓	✓	
Statistical	MAP	(Rasmussen and Hager, 2001)	D	✓	✓	✓		✓	✓	X
		(Cremers and Soatto, 2005)	D	✓	✓	✓		✓	RA	X
		(Shen et al, 2007)	D	✓	✓	✓	✓	✓	✓	CX
	PF	(Vaswani et al, 2007)	D	✓	✓	✓		✓	RN	X
	EM	(Stolkin et al, 2008)	D	✓	✓	✓	✓	✓	R	CX
Wavelets		(Wiskott, 1997)	F		✓			✓	R	
		(Kong et al, 1998)	F	✓	✓		✓	✓	R	X
O.F.		(Zhang et al, 2007)	F		✓			I	RA	C
		(Xu et al, 2008)	D	✓	✓	✓		I	✓	
		(Klappstein et al, 2009)	F	✓	✓			I	RA	X
		(Bugeau and Pérez, 2009)	F/D	✓	✓		✓	I	R	
		(Ommer et al, 2009)	F	✓	✓			I	R	
Layers		(Kumar et al, 2008)	F	✓	✓	✓	✓	✓	RA	X
		(Min and Medioni, 2008)	D	✓	✓	✓	✓	✓	✓	X
Manifold Clustering	Iter	(Fischler and Bolles, 1981)	F			✓	✓	I	RA	C
		(Ho et al, 2003)	F			✓		✓	✓	CD
		(da Silva and Costeira, 2008)	F			✓		✓	✓	X
	Stat	(Kanatani and Matsunaga, 2002)	F			✓		I	R	
		(Sugaya and Kanatani, 2004)	F			✓		I	R	C
		(Gruber and Weiss, 2004b)	F			✓	✓	I	R	X
		(Gruber and Weiss, 2006)	F	✓	✓	✓	✓	I	R	X
	ALC	(Rao et al, 2008)	F	✓		✓	✓	I	R	
	Fact	(Costeira and Kanade, 1998)	F			✓		I	R	
		(Ichimura and Tomita, 2000)	F			✓		I	R	
		(Zelnik-Manor and Irani, 2003)	F			✓		✓	RA	CD
		(Zhou and Huang, 2003)	F	✓		✓	✓	✓	RA	CD
	Subspaces	(Vidal and Hartley, 2004)	F	✓		✓		✓	R	C
		(Yan and Pollefeys, 2006, 2008)	F			✓		✓	✓	CDX
		(Julia et al, 2008)	F	✓		✓		✓	R	C
		(Goh and Vidal, 2007)	F			✓		✓	R	CD
		(Vidal et al, 2008)	F	✓		✓		✓	R	C
		(Goh and Vidal, 2008)	F			✓		✓	R	CD
		(Chen and Lerman, 2009)	F			✓		✓	✓	CD
		(Zappella et al, 2009)	F			✓	✓	✓	✓	C
Features (F) / Dense (D)										
Occlusion or Missing Data										
Spatial Continuity										
Temporary Stopping										
Robustness (Noise, Outliers, Initialization)										
Dependency (I independent, D dependent, ✓all)										
Kind (R rigid, N non-rigid, A articulated, ✓all)										
Prior knowledge (C Clusters number, D Subspace dimension, X Other, T Training)										

Table 1. Summary of the examined techniques with respect to the most important attributes.

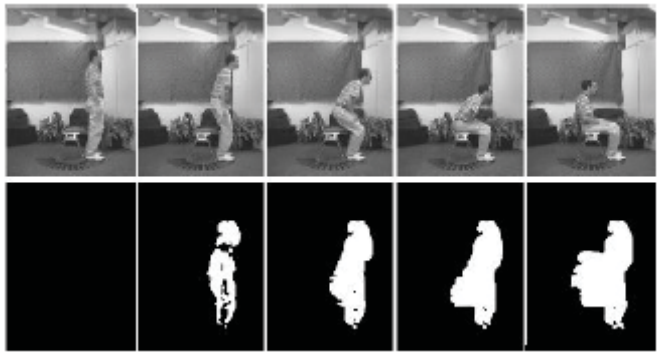

Fig. 1. Example of an image difference result (Bobick and Davis, 1996).

problems of image difference techniques such as dealing with temporary stopping. Another example is (Colombari et al, 2007), where in order to deal with noise and very small camera movements the authors propose a robust statistic to model the background.
As can be seen from the table 1, image difference is mainly based on dense representation of the objects. It combines simplicity and good overall results being able to deal with occlusions, multiple objects, independent motions, non-rigid and articulated objects. The main problem of these techniques is the difficulty to deal with temporary stopping and with moving cameras. In order to be successful in these situations a history model of the background needs to be built. Furthermore, image difference algorithms are still very sensitive to noise and to light changes, hence they cannot be considered an ideal choice in case of cluttered background.

2.2.2 Statistical framework

Statistical theory is widely used in the motion segmentation field. In fact, motion segmentation, in the simple case, can be seen as a classification problem where each pixel has to be classified as background or foreground. Statistical approaches can be further divided depending on the framework used. Common frameworks are Maximum A posteriori Probability (MAP), Particle Filter (PF) and Expectation Maximization (EM). Statistical approaches provide a general tool that can be used in very different ways depending on the specific technique.
In (Rasmussen and Hager, 2001), a MAP framework is used, namely they use the Kalman Filter and the Probabilistic Data Association Filter, to predict the most likely location of a known target in order to initialize the segmentation process. Another technique based on MAP is (Cremers and Soatto, 2005), where level sets (Sethian, 1998) are used to incorporate motion information. In (Shen et al, 2007), MAP formulation is proposed to iteratively update the motion fields and the segmentation fields along with the high-resolution image. The formulation is solved by a cyclic coordinate descent process that treats motion, segmentation, and high-resolution image as unknowns, and estimates them jointly using the available data. Another widely used statistical framework is PF. The main aim of PF is to track the evolution of a variable over time. The basis of the method is to construct a sample-based representation of the probability density function. Basically, a series of actions are taken, each of them modifying the state of the variable according to some model. Multiple

copies of the variable state (particles) are kept, each one with a weight that signifies the quality of that specific particle. An estimation of the variable can be computed as a weighted sum of all the particles. The PF algorithm is iterative and each iteration is composed by prediction and update. After each action particles are modified according to the model (prediction), then each particle weight is re-evaluated according to the information extracted from an observation (update). At every iteration, particles with small weights are eliminated (Rekleitis, 2003). An example of PF applied to motion segmentation is (Vaswani et al, 2007), where some well known algorithms for object segmentation using spatial information, such as geometric active contours (Blake, 1999) and level sets (Sethian, 1998), are unified using a PF framework.

EM is also a frequently exploited framework in motion segmentation. The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in presence of missing or hidden data. In ML the aim is to estimate the model parameter(s) for which the observed data is most likely to belongs to. Each iteration of the EM algorithm consists of an E-step and an M-step. In the E-step, using the conditional expectation the missing data are estimated. Whereas, in the M-step the likelihood function is maximized. Convergence is assured since the algorithm is guaranteed to increase the likelihood at each iteration (Borman, 2004). An example of EM applied to motion segmentation is (Stolkin et al, 2008), where the authors present an algorithm which uses EM and Extended-Markov Random Field (E-MRF). In order to track the camera trajectory (egomotion), the algorithm merges the observed data (the current image) with the prediction derived from prior knowledge of the object being viewed. The merging step is driven by the E-MRFs within a statistical framework.

Statistical approaches use mainly dense based representation. They work well with multiple objects and can deal with occlusions and temporary stopping. In general they are robust as long as the model reflects the actual situation but they degrade quickly as the model fails to represent the reality. Moreover, most of the statistic approaches require some kind of a priori knowledge.

2.2.3 Wavelets

Another group of motion segmentation algorithms is based on wavelets analysis. These methods exploit the ability of wavelets to perform analysis of the different frequency components of the images, and then study each component with a resolution matched to its scale. Usually wavelet multi-scale decomposition is used in order to reduce the noise and in conjunction with other approaches, such as optical flow, applied at different scales. However, there are a few proposals where wavelet is the main segmentation algorithm. In (Wiskott, 1997) the author combines Gabor and Mallat wavelet transform to overcome the aperture problem and the correspondence problem. The former transform is used to estimate the motion field and roughly cluster the image, while the latter is used to refine the clustering. The main limitation of this model is that it assumes that the objects only translate in front of the camera. A different approach is presented in (Kong et al, 1998) where the motion segmentation algorithm is based on Galilean wavelets. These wavelets behave as matched filters and perform minimum mean-squared error estimations of velocity, orientation, scale and spatio-temporal positions. This information is finally used for tracking and segmenting the objects.

Fig. 2. Example of OF, in red the vectors of the flow of the moving person

Wavelets solutions seem to provide overall good results but limited to simple cases (such as translation in front of the camera). Wavelets were in fashion during the 90s, nowadays the research interest seems to be less active, at least in relation to motion segmentation.

2.2.4 Optical flow

Optical flow (OF) can be defined as the apparent motion of image brightness patterns in an image sequence. An example of OF can be seen in figure 2. Like image difference, OF is an old concept greatly exploited in computer vision. It was first formalized and computed for image sequences by Horn and Schunck in the 1980 (Horn and Schunck, 1980). However, the idea of using discontinuities in the optical flow in order to segment moving objects is even older, in (Horn and Schunck, 1980) there is a list of older methods based on this idea but they all assume the optical flow is already known. Since the work of Horn and Schunck, many other approaches have been proposed. In the past, the main limitation of such methods was the high sensitivity to noise and the high computational cost. Until recently, OF was more often used in hardware implementations in order to overcome the computational cost, as in (Jos et al, 2005). Nowadays, thanks to the high computational speed and to improvements made by research, OF is widely used also in software implementation. In (Xu et al, 2008) is presented a variational formulation of OF combined with color segmentation obtained by the Mean-shift algorithm. The authors of (Klappstein et al, 2009) exploit OF in order to built a robust obstacle detection for driver assistance purposes. The work is done both with monocular (exploiting some motion constraints) and stereo (using Extended Kalman Filter) vision. In (Bugeau and Pérez, 2009) the segmentation problem is addressed by combining motion information, spatial continuity and photometric information. In (Ommer et al, 2009) an algorithm for segmentation, tracking and object recognition is presented. The segmentation and tracking parts are done by OF (using salient features and KLT tracking). The algorithm is based on grouping together salient features following a proximity criteria. The features are tracked by KLT and the mean flow is computed. The position of the group of features is predicted using the previous mean flow in order to constrain the tracking area. At every iteration the mean flow is updated taking into account the old flows with an exponential decay over time.
OF is, theoretically, a good clue in order to segment motion. However, alone it is not enough because it cannot deal with occlusions and temporal stopping. Statistical techniques or

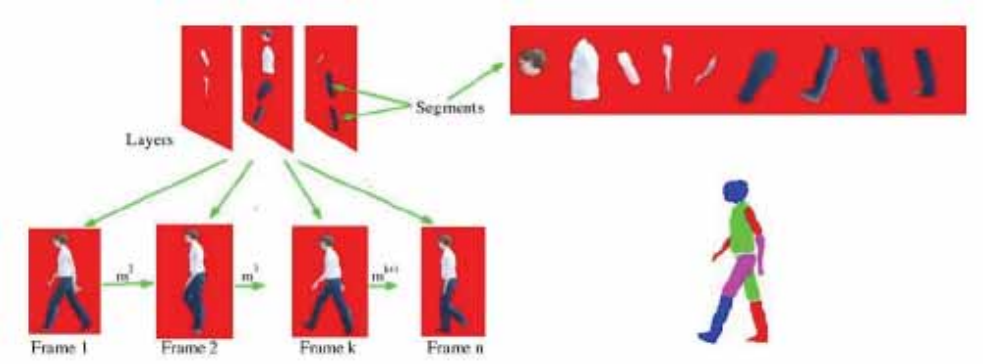

Fig. 3. Example of layers segmentation (Kumar at al, 2008)

spatial analysis (like colour or texture) could help to increase the robustness as OF is still very sensitive to noise and light changes.

2.2.5 Layers

The first layers technique was proposed by J. Wang and E. Adelson in 1993 (Wang and Adelson, 1993). The key idea of layers based techniques is to divide the image into layers with uniform motion. Furthermore, each layer is associated with a depth level and a "transparency" level that determines the behaviour of the layers in case of overlapping. This approach is often used in stereo vision as the depth distance can be recovered easily. However, even without computing the depth it is possible to estimate which objects move on similar planes. This is extremely useful as it helps to solve the occlusion problem. Recently, new interest raised around this idea (Kumar et al, 2008; Min and Medioni, 2008). The authors of (Kumar et al, 2008) propose a method for learning a layered representation of the scene. They initialize the algorithm by first finding coarse moving components between every pair of frames. They divide the image in patches and find the rigid transformation that moved the patch from frame j to frame j + 1. The initial estimate is then refined using αβ-swap and α-expansion algorithms (Boykov et al, 1999). More recently, in (Min and Medioni, 2008), a new layer based technique was presented. This technique exploits a 5 dimensional representation of each feature, the 5D token is composed by: position (x, y), time (t), and velocity (v_x, v_y). The layers are seen as 3D variety which can be extracted from the 5D tensor (using neighbours tokens) by tensor voting framework. In order to produce accurate results pre-segmented areas based on color segmentation (performed by Mean-shift) are required. An example of a layer segmentation is shown in figure 3.

Layers solutions are very interesting. It is probably the most natural solution for the occlusion problem: human beings also use depth perception to solve this issue. The main drawback is the level of complexity of these algorithms and the number of parameters that have to be tuned manually. Furthermore, a deeper evaluation should be carried out as none of the presented algorithms has exhaustive tests with more than two motions.

2.2.6 Manifold clustering

Manifold clustering techniques consist in projecting the original data into a smaller space (if necessary, otherwise the ambient space could be directly used) and trying to cluster together data that has common properties by, for example, fitting a set of hyperplanes to the data. Nowadays, manifold clustering is a "hot" topic and it is applied in many fields. Segmentation seems one of the most natural applications, particularly motion segmentation. This class of solutions is usually based on feature points. They provide not only the segmentation but they can be naturally extended to Structure from Motion (SfM) in order to recover the 3D structure of the objects and the motion of the camera. Furthermore, they do not have any problem with temporary stopping because features can be tracked even if the object is not moving (provided that this is a temporary situation). Most of these techniques assume an affine camera model, however, it is possible to extend them to the projective case by an iterative process as shown in (Li et al, 2007). A common drawback to all these techniques is that they can deal very well when the assumptions of rigid, independent and non degenerate motions, are respected, but if one of these assumptions fails, then problems start to arise as the properties of motions have to be taken into account explicitly. This group of techniques is rather big, hence, a further classification helps to give some order. Manifold clustering can be divided into, Iterative solutions, Statistical solutions (solutions that fall inside this category could be placed in the previous Statistical group, but in this case we refer to statistical frameworks specifically applied to manifold clustering), Agglomerate Lossy Compression (ALC), Factorization solutions and Subspace Estimation solutions.

An iterative solution is presented in (Fischler and Bolles, 1981) where the RANdom SAmple Consensus (RANSAC) algorithm is used. RANSAC tries to fit a model to the data randomly sampling n points, then it computes the residual of each point to the model and those points whose residual is below a threshold are considered inliers. The procedure is repeated until the number of inliers is above a threshold, or enough samples have been drawn. Another iterative algorithm called "K-Subpsace Clustering" is presented in (Ho et al, 2003) for face clustering, however, the same idea could be adopted to solve the motion segmentation problem. K-Subpsace can be seen as a variant of K-means. K-Subspace iteratively assigns points to the nearest subspace, than that subspace is updated computing the new bases that minimize the sum of the square distances to all the points of that cluster. The algorithm ends after a predefined number of iterations. The authors of (da Silva and Costeira, 2008) propose a subspace segmentation algorithm based on a Grassmannian minimization approach. The technique consists in estimating the subspace with the maximum consensus (MCS): maximum number of data that are inside the subspace. Then, the algorithm is recursively applied to the data inside the subspace in order to look for smaller subspaces included in it.

Iterative approaches are in general robust to noise and outliers, and they provide good solutions if the number of clusters and the dimension of the subspaces are known. This prior knowledge can be clearly seen as their limitation as this information is not always available. Moreover, they require an initial estimation and they are not robust against bad initializations, so when the initialization is not close enough to the correct solution the algorithms are not guaranteed to converge.

Another manifold clustering group is composed by statistical solutions. In (Kanatani and Matsunaga, 2002) the authors use a statistical framework for detecting degeneracies of a geometric model. They use the geometric information criterion (AIC) defined in (ichi Kanatani, 1997) in order to evaluate whether two clouds of points should be merged or not.

Another statistical based technique is (Sugaya and Kanatani, 2004). This paper analyses the geometric structure of the degeneracy of the motion model, and suggests a multi-stage unsupervised learning scheme first using the degenerate motion model and then using the general 3-D motion model. The authors of (Gruber and Weiss, 2004b) extend the EM algorithm already proposed in (Gruber and Weiss, 2004a) for the single object case in order to deal with multiple objects and missing data. In (Gruber and Weiss, 2006) the same authors further extend the method incorporating non-motion cues (such as spatial coherence) into the M step of the algorithm.

Statistical solutions have more or less the same strengths and weaknesses of iterative techniques. They can be robust against noise whenever the statistical model is built taking the noise explicitly into account. However, when noise is not considered or is not modeled properly their performances degenerate rapidly. As previously said for general statistical approaches: they are robust as long as the model reflects the actual situation.

A completely different idea is the basis of (Rao et al, 2008) which uses the Agglomerative Lossy Compression (ALC) algorithm (Ma et al, 2007). This technique consists in minimizing a cost function by grouping together trajectories. Roughly speaking, the cost function is given by the amount of information required to represent each manifold given a particular segmentation.

ALC provides a connection between coding theory and space representation. It performs extremely well with a variety of motions. However, it has some problems to deal with a lot of data (curse of dimensionality). Furthermore, the algorithm depends on a parameter that has to be tuned per each sequence depending on the number of motions and the amount of noise. Although the tuning can be automated trying many different values and choosing at the end the solution with the lowest cost, this process is highly time-consuming.

Factorization techniques are based on the approach introduced by Tomasi and Kanade in 1992 (Tomasi and Kanade, 1992) to recover structure and motion using features tracked through a sequence of images. In (Costeira and Kanade, 1998) the framework of Tomasi and Kanade is first used in order to factorize the trajectory matrix W by Singular Value Decomposition into the matrices U, D, and V. Then a matrix called "shape interaction matrix" $Q = VV^T$ is built. The shape interaction matrix has, among other properties, zero entries if the two indexes represent features belonging to different objects, non-zero otherwise. Hence, the algorithm focuses on finding the permutation of the interaction matrix that gives a block diagonal matrix structure as shown in figure 4. In (Ichimura and Tomita, 2000), once the rank r of the trajectory matrix is estimated they perform the QR decomposition of the shape interaction matrix and select the r bases of the shape space which gives the segmentation among those features. Finally, the remaining features are segmented by using the orthogonal projection matrix. The two previous factorization techniques assume that the objects have independent motions. In (Zelnik-Manor and Irani, 2003) the authors study the degeneracy in case of dependent motion. They propose a factorization method that consists in building an affinity matrix by using only the dominant eigenvector and estimating the rank of the trajectory matrix by studying the ratio between the eigenvalues. In (Zhou and Huang, 2003) a hierarchical factorization method for recovering articulated hand motion under weak perspective projection is presented. They consider each part of the articulated object as independent and they use any of the techniques able to deal with missing data to fill the gaps. In the second step, they guarantee that the end of the consecutive objects are linked in the recovered motion.

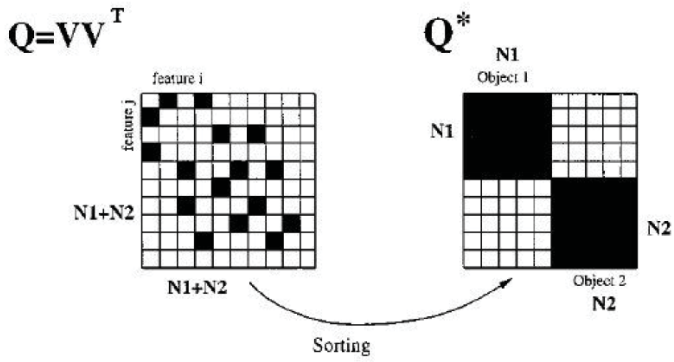

Fig. 4. (Costeira and Kanade, 1998) computes the interaction matrix Q and finds the permutation of rows and columns that gives a block diagonal matrix.

Factorization techniques are based on a very simple and elegant framework. However, factorization methods are particularly sensitive to noise and cannot deal very well with outliers. Moreover, most of these techniques assume rigid and independent motions.
The last category of manifold clustering is the subspace estimation techniques.
The work presented in (Vidal and Hartley, 2004) belongs to this group. First, exploiting the fact that trajectories of rigid and independent motion generate subspaces at most of dimension four, they project the trajectories onto a five dimensional space using PowerFactorization. Then, the Generalized Principal Component Analysis (GPCA) is used 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. More recently, the same authors in (Vidal et al, 2008) extended the previous explained framework using RANSAC to perform the space projection in order to be able to deal with outliers. Another well known technique is the Local Subspace Affinty (LSA) (Yan and Pollefeys, 2006, 2008). LSA is able to deal with different types of motion: independent, articulated, rigid, non-rigid, degenerate and non-degenerate. The key idea is that different motion trajectories lie in subspaces of different dimension. Thus, the subspaces are estimated and an affinity matrix is built using principal angles. The final segmentation is obtained by clustering the affinity matrix. The main limitations of LSA are the difficulty of estimating the size of the global and local subspaces without manual tuning, and the fact that a full trajectory matrix without missing data is assumed. In (Julia et al, 2008) a technique similar to LSA is presented in order to deal with missing data. The idea is to fill the missing data using a frequency spectra representation for the matrix estimation. When a full trajectory matrix is obtained an affinity matrix is built and a cluster algorithm based on normalized cuts is applied in order to provide the segmentation. In (Chen and Lerman, 2009) the authors propose a generalization
of LSA called Spectral Curvature Clustering (SCC). SCC differs from LSA for two main reasons. The first reason is related to the affinity measure, SCC uses polar curvature while LSA uses principal angles. In SCC the affinity between a point i and the other points is given by the polar curvature of the space generated by i and some combination of other d + 1 points (where d is the size of the generated subspace). The second reason is how they select

which points have to be combined with I: SCC uses an iterative solution while LSA uses a nearest neighbour solution. Theoretically, in SCC all the possible combination of points should be tried but this may not be computationally feasible, instead, only one combination of d + 1 points is randomly selected among the points that belong to the same cluster of i. Of course, the first time this selection is done, there is no information about which point belong to which cluster, hence at the first iteration the points are randomly selected among all of them. At the second iteration, the clustering result of the first iteration is used to constrain the selection among the points that were clustered with i. A completely different strategy is presented in (Goh and Vidal, 2007) where, starting from the Locally Linear Embedding algorithm (Saul and Roweis, 2003), they propose the Locally Linear Manifold Clustering Algorithm (LLMC). With LLMC the authors try to deal with linear and non-linear manifolds. The same authors extended this idea to Riemannian manifolds (Goh and Vidal, 2008). They project the data from the Euclidean space to a Riemannian space and reduce the clustering to a central clustering problem. Finally, in (Zappella et al, 2009) the authors enforce the LSA algorithm proposing a new Enhanced Model Selection (EMS) technique. EMS is a generic rank estimation tool, in this case it is used in order to estimate the size of the global and local subspaces in an automatic fashion, auto-tuning the parameters in order to deal with different noise conditions and different number of motions.
Subspace estimation techniques can deal with intersection of the subspaces and generally they do not need any initialization. However, all these techniques suffer from common problems: curse of dimensionality, weak estimations of number of motions and subspaces dimension. The curse of dimensionality is mainly solved in two ways: projection into smaller subspaces or random sampling. Whereas the number of motions and the subspace dimension estimations are commonly two open issues.

3. Discussions and conclusions

Table 2 summarises and generalises the advantages and disadvantages of each group of techniques. This review should have given an idea of how vast the motion segmentation literature is, and the fact that research in this field is still active (most of the papers presented here were published after 2005) is a sign of the importance of this problem. On the other hand, effervescent research activity signifies also that many problems have still to be solved and there is not an outstanding solution yet. From the analysis it is possible to state that manifold clustering algorithms seems one of the most natural solutions for motion segmentation. Recently manifold clustering has been studied and exploited deeply in order to solve the motion segmentation problem. This class of techniques have already good performances, nevertheless there is space for further improvements. A quick glance at table 1 may catch the attention on the fact that for manifold clustering techniques, the price to pay in order to be able to deal with different kind of motions and with dependent motions is a higher amount of prior knowledge (in particular about the dimension of the generated subspaces). The amount of prior knowledge is another limitation that in future should be overcome. In order to obtain more robust results it would be interesting to study different ways of merging spatial information, and to exploit the ability of statistical frameworks to find hidden information and outliers.

Techniques		Pros		Cons	
Image Diff.		- Simple - Occlusions	- Dependency - Kind of motion	- Sensitive to noise - Temporary stopping	- Moving camera
Statistical		- Occlusions - Temporary stopping		- Sensitive to model - Prior knowledge	- Dependency
Wavelets		- Depth estimation	- Dependency	- Multiple motions	- Kind of motion
O.F.		- Simple		- Dependency - Non-rigid motions	- Sensitive to noise
Layers		- Occlusions		- Complexity	- Many Parameters
Manifold Clustering	Iter	- Extension to SfM - Temporary stopping		- Prior knowledge - Sensitive to initialization - Kind of motion	- Occlusions - Dependency
	Stat	- Extension to SfM - Temporary stopping		- Prior knowledge - Dependency - Kind of motion	- Occlusions - Sensitive to model
	ALC	- Extension to SfM - Temporary stopping	- Occlusions - Misclassification	- Time consuming - Dependency	- No justification - Curse of dimensionality
	Fact	- Extension to SfM - Temporary stopping	- Elegant	- Prior knowledge - Dependency - Kind of motion	- Occlusions - Sensitive to noise
	Sub	- Extension to SfM - Temporary stopping	- Misclassification	- Prior knowledge - Curse of dimensionality	- Occlusions

Table 2. Summary and generalisation of pros and cons of each group of techniques.

Nowadays the misclassification rates knowing the number of motions are already quite good. Despite the fact that the misclassification rates could be further improved, it is the opinion of the authors that future works should focus on the ability to estimate the number of clusters in a more efficient way. In general feature based techniques are preferred over dense based approaches as the amount of computation required by dense approaches is very large. However, feature based techniques have to rely on the ability of the tracker to find salient points and track them successfully through the video sequence. Today, such an assumption is not too constraining but it is important to develop algorithms able to deal only with few points (from four to six) per motion instead of requiring lots of them. Moreover, in order to have a useful system for real time applications, future motion segmentation algorithms should be able to work incrementally. An ideal incremental algorithm should be able to refine the segmentation at every new frame (or every group of few frames) without recomputing the whole solution from the beginning.

4. Acknowledgements

This work has been supported by the Spanish Ministry of Science projects DPI2007-66796-C03-02 and DPI2008-06548-C03-03/DPI. L. Zappella is supported by the Catalan government scholarship 2007FI_A 00765.

5. References

Blake A (1999) Active contours. Robotica 17(4):459–462

Bobick A, Davis J (1996) An appearance-based representation of action. IEEE InternationalConference on Pattern Recognition pp 307–312

Borman S (2004) The expectation maximization algorithm – a short tutorial

Boykov Y, Veksler O, Zabih R (1999) Fast approximate energy minimization via graph cuts. In: International Conference on Computer Vision, pp 377–384

Bugeau A, Perez P (2009) Detection and segmentation of moving objects in complex scenes. Computer Vision and Image Understanding 113:459–476

Cavallaro A, Steiger O, Ebrahimi T (2005) Tracking Video Objects in Cluttered Background. IEEE Transactions on Circuits and Systems for Video Technology 15(4):575–584

Chen G, Lerman G (2009) Spectral curvature clustering (scc). International Journal of Computer Vision 81:317–330

Cheng FH, Chen YL (2006) Real time multiple objects tracking and identification based on discrete wavelet transform. Pattern Recognition 39(6):1126–1139

Colombari A, Fusiello A, Murino V (2007) Segmentation and tracking of multiple video objects. Pattern Recognition 40(4):1307–1317

Costeira JP, Kanade T (1998) A multibody factorization method for independently moving objects. International Journal of Computer Vision 29(3):159–179

Cremers D, Soatto S (2005) Motion competition: A variational approach to piecewise parametric motion segmentation. International Journal of Computer Vision 62(3):249–265

Fischler MA, Bolles RC (1981) Ransac random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24:381–395

Goh A, Vidal R (2007) Segmenting motions of different types by unsupervised manifoldclustering. IEEE Conference on Computer Vision and Pattern Recognition pp 1–6

Goh A, Vidal R (2008) Clustering and dimensionality reduction on riemannian manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition

Gruber A, Weiss Y (2004a) Factorization with uncertainty and missing data: exploiting temporal coherence. Advances in Neural Information Processing Systems

Gruber A, Weiss Y (2004b) Multibody factorization with uncertainty and missingdata using the em algorithm. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1:707–714

Gruber A, Weiss Y (2006) Incorporating non-motion cues into 3d motion segmentation. In: European Conference on Computer Vision, pp 84–97

Ho J, Yang MH, Lim J, Lee KC, Kriegman D (2003) Clustering appearances of objects under varying illumination conditions. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol 1, pp 11–18

Horn BK, Schunck BG (1980) Determining optical flow. Tech. rep., Cambridge, MA, USA

Ichimura N, Tomita F (2000) Motion segmentation based on feature selection from shape matrix. Systems and Computers in Japan 31(4):32–42

Jos LM, Zuloaga A, Cuadrado C, Lzaro J, Bidarte U (2005) Hardware implementation of optical flow constraint equation using fpgas. Computer Vision and Image Understanding 98(3):462–490

Julia C, Sappa A, Lumbreras F, Serrat J, Lopez A (2008) Rank estimation in 3d multibody motion segmentation. Electronics Letters 44(4):279–280

Ichi Kanatani K (1997) Statistical optimization and geometric visual inference. In: AFPAC '97: Proceedings of the International Workshop on Algebraic Frames for the Perception-Action Cycle, Springer-Verlag, pp 306–322

Kanatani K, Matsunaga C (2002) Estimating the number of independent motions for multibody motion segmentation. In: Proceedings of the Fifth Asian Conference on Computer Vision, vol 1, pp 7–12

Klappstein J, Vaudrey T, Rabe1 C, Wedel A, Klette R (2009) Moving object segmentation using optical flow and depth information. In: Pacific-Rim Symposium on Image and Video Technology, p 611623

Kong M, Leduc JP, Ghosh B, Wickerhauser V (1998) Spatio-temporal continuous wavelet transforms for motion-based segmentation in real image sequences. Proceedings of the International Conference on Image Processing 2:662–666

Koterba S, Baker S, Matthews I, Hu C, Xiao J, Cohn JF, Kanade T (2005) Multi-view aam fitting and camera calibration. In: ICCV, pp 511–518

Kumar MP, Torr PH, Zisserman A (2008) Learning layered motion segmentations of video. International Journal of Computer Vision 76(3):301–319

Li R, Yu S, Yang X (Aug. 2007) Efficient spatio-temporal segmentation for extracting moving objects in video sequences. IEEE Transactions on Consumer Electronics 53(3):1161–1167

Li T, Kallem V, Singaraju D, Vidal R (2007) Projective factorization of multiple rigid-body motions. pp 1–6

Llado X, Bue AD, Agapito L (2006) Euclidean reconstruction of deformable structure using a perspective camera with varying intrinsic parameters. 18th International Conference on Pattern Recognition 1:139–142

Ma Y, Derksen H, Hong W, Wright J (2007) Segmentation of multivariate mixed data via lossy data coding and compression. IEEE transactions on pattern analysis and machine intelligence 29(9):1546 –1562

Min C, Medioni G (2008) Inferring segmented dense motion layers using 5d tensor voting. IEEE transactions on pattern analysis and machine intelligence 30(9):1589–1602

Ommer B, Mader T, Buhmann JM (2009) Seeing the objects behind the dots: Recognition in videos from a moving camera. International Journal of Computer Vision 83:57–71

Rao SR, Tron R, Vidal R, Ma Y (2008) Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition

Rasmussen C, Hager GD (2001) Probabilistic data association methods for tracking complex visual objects. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(6):560–576

Rekleitis I (2003) Cooperative localization and multi-robot exploration. PhD in computer science, School of Computer Science, McGill University, Montreal, Quebec, Canada

Saul LK, Roweis ST (2003) Think globally, fit locally: unsupervised learning of low dimensional manifolds. J Mach Learn Res 4:119–155

Sethian J (1998) Level set methods and fast marching methods: Evolving interfaces in computational geometry

Shen H, Zhang L, Huang B, Li P (2007) A map approach for joint motion estimation, segmentation, and super resolution. IEEE Transactions on Image Processing 16(2):479–490

da Silva NP, Costeira JP (2008) Subspace segmentation with outliers: a grassmannian approach to the maximum consensus subspace. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1–6

Stolkin R, Greig A, Hodgetts M, Gilby J (2008) An em/e-mrf algorithm for adaptive model based tracking in extremely poor visibility. Image and Vision Computing 26(4):480–495

Sugaya Y, Kanatani K (2004) Geometric structure of degeneracy for multi-body motion segmentation. In: Statistical Methods in Video Processing, pp 13–25

Tomasi C, Kanade T (1992) Shape and motion from image streams under orthography: a factorization method. International Journal of Computer Vision 9(2):137–154

Vaswani N, Tannenbaum A, Yezzi A (2007) Tracking deforming objects using particle filtering for geometric active contours. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(8):1470–1475

Vidal R, Hartley R (2004) Motion segmentation with missing data using powerfactorizationand gpca. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2:310–316

Vidal R, Tron R, Hartley R (2008) Multiframe motion segmentation with missing data using powerfactorization and gpca. International Journal of Computer Vision 79:85–105

Wang J, Adelson E (1993) Layered representation for motion analysis. pp 361–366

Wiskott L (1997) Segmentation from motion: Combining Gabor- and Mallat-wavelets to overcome aperture and correspondence problem. In: Sommer G, Daniilidis K, Pauli J (eds) Proceedings of the 7th International Conference on Computer Analysis of Images and Patterns, Springer-Verlag, Heidelberg, vol 1296, pp 329–336

Xu L, Chen J, Jia J (2008) A segmentation based variational model for accurate optical flow estimation. In: European Conference on Computer Vision, pp 671–684

Yan J, Pollefeys M (2006) A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In: Computer Vision ECCV 2006, vol 3954, pp 94–106

Yan J, Pollefeys M (2008) A factorization-based approach for articulated nonrigid shape, motion and kinematic chain recovery from video. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(5):865–877

Zappella L, Llado X, Salvi J (2009) Rank estimation of trajectory matrix in motion segmentation. Electronics Letters 45(11):540–541

Zelnik-Manor L, Irani M (2003) Degeneracies, dependencies and their implications in multi-body and multi-sequence factorizations. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2:287–93

Zhang J, Shi F, Wang J, Liu Y (2007) 3d motion segmentation from straight-line optical flow. In: Multimedia Content Analysis and Mining, pp 85–94

Zhou H, Huang TS (2003) Recovering articulated motion with a hierarchical factorization method. In: Gesture Workshop, pp 140–151