Tracking powered by Superpixels
Reviews of Papers

Ahmad Humayun
Dept. of Computer Science,
University College London
ahmad.humayun.09@ucl.ac.uk

1 Heitz and Bouthemy [1993]

1. Solves dense motion estimation problem by solving “several complementary constraints” - optical flow estimation while preserving motion boundaries. It mainly takes care of discontinuities and occlusions in motion fields.

2. Theoretical framework relies on Bayesian estimation using MRF as a global statistical model.

3. Aims:

   (a) Optical flow while preserving motion boundaries

   (b) Processing occlusion regions

   (c) Fusion of gradient and feature based constraints for local motion measurement.

4. Multiple motion constraints:

   (a) Local and Global optimisation techniques to solve the aperture problem. But all of these **gradient based** techniques have problems in occluded regions, motion discontinuities, intensity discontinuities, smooth regions, linear variation regions and wherever there is large motion.

   “Wohn and Waxman, [35] study the global analytic structure of a 2-D motion field and propose a segmentation method based on the recovery of boundaries between regions of analycity in the optical flow field.”

   (b) Gradient based motion constraint used is derived from our multimodal motion estimator (similar to Horn and Schunck [18]) by retaining only the image flow constraint and discarding motion discontinuities and occlusion areas. Poor estimates are observed both on the central bright intensity lines and in the occlusion areas like in all other standard smoothing methods.

   (c) The **moving edge** estimation is done by thresholding a log-likelihood ratio test based on parameters defining surface S - where the parameters are tuned to detect edge locations, orientation and displacement.

   (d) Both constraints need to pass some validation. In case of moving edge, thresholding the log-likelihood itself brings a natural way to validate an edge. For checking image flow constraint, consistency of spatial derivatives across time is checked. This is done by thresholding the log-likelihood ratio of the hypothesis that the spatial derivative is the same under some Gaussian noise.

5. “Global Bayesian estimation defines a coherent mathematical framework to extract labels describing motion from image sequences.”:

   (a) use image features as observations in the estimation process.

   (b) these observations are combined to derive estimates of the unknown labels

6. the relationship between the observation fields and motion labels is specified using MRF: The following things in the framework of motion estimation are used to find the MAP estimate:

   (a) regularisation of the velocity field along with preservation of motion boundaries

   (b) multimodal cooperation between different measurement sources

   (c) discarding of invalid local motion constraints (in particular in the occlusion regions)

   (d) processing of motion discontinuities.

   [6,25,27,34,35]

2 Black and Jepson [1996]

1. Estimates optical flow through motion of planar regions and local deformations (deformations allowed since the assumption of planarity is likely to be violated)

2. Segmentation is done on brightness values to constrain motion to planar regions (those which will have coherent motion) in the scene (helps finding motion boundaries precisely). It uses an analog spatial outlier process to define discontinuity between pixels - also a penalty term is defined which needs to be paid with increasing discontinuity - eventually to minimise an objective function with a data term and spatial coherence term. [13]

3. Algorithm can be broken to these coarse levels:

   (a) First step: estimate a coarse fit to the parametric flow model and estimate some set of parameters for the region:

      i. Low level smoothing image brightness while checking for discontinuities

      ii. Another low level process providing estimates of motion

   (b) Second step: Refine the initial fit with "standard area-based regression approaches". This medium level tries to collate low level information by connecting piecewise smooth brightness regions and then estimating their motion. This is done by:

      i. Fits parametric model to the coarse motion estimate, to give initial estimate of motion. A translational/affine/planar model is found which best captures image motion in the region.

      ii. Using the model, the region is warped into alignment. Gradient-based optical flow are computed for these warped regions and help in refining the initial parametric model.

      iii. The low-level patches are allowed to deform by using optical flow constraints, spatial coherence, and the motion estimate of the planar patch

   [1,22,35,46]

3 Chang et al. [1997]

1. The motion field is taken as a sum of a parametric field and a non-parametric residual field.

2. The parameters for the parametric field are found by a least squares solution given the best estimates of the motion field and segmentation.

3. "The motion field is refined by estimating the minimum-norm residual field given the best estimate of the parametric field, under the constraint that motion field be smooth within each segment."

4. "The segmentation field is refined to yield the minimum-norm residual field given the best estimate of the motion field, using Gibbsian priors." (Gibbsian priors are used to encourage connectivity among the segmentation labels to avoid small, isolated regions.)

5. The interdependence between the optical flow field and the segmentation map is imposed through the Bayesian framework.

6. The least squares estimates of the mapping parameters Φ for each segmentation class K can be computed in closed-form given the MAP estimate:

   $$(\hat{u}, \hat{v}, \hat{x}) = \max_{u,v,x} P(g_k|u, v, x, g_{k-1})P(u, v|x, g_{k-1})P(x|g_{k-1})$$

   where u, v are motion field vectors, the two frames are g_k (current frame), g_{k-1} (search frame), and x is the segmentation field. The conditional PDF $P(g_k|u, v, x, g_{k-1})$ quantifies how well the motion and segmentation estimates fit the given frames. This PDF is modelled by a Gibbs distribution.

   $P(u, v|x, g_{k-1})$ is also modelled by a Gibbs distribution with a potential function which aims to minimise the deviation of the motion field from u, v from the parametric motion u_p, v_p. Note here that u_p, v_p

are functions of segmentation labels x and the mapping parameters Φ (which in turn is a function of $u, v,$ and x as given above.

The third term $P(x|g_{k-1})$, the priori probability of the segmentation follows a Gibbs distribution to encourage formation of contiguous regions.

7. The dense representation of the residual field has been proposed for improved motion segmentation.

4 Shi and Malik [1998]

1. Form a graph $G = (V, E)$ where the nodes are pixels connected with other pixels in its spatiotemporal neighbourhood and the weight $w(i, j)$ is given by the similarity (colour, brightness, texture, motion, disparity, etc. - the paper only uses motion) between the pixel nodes.
2. Motion similarity is measured at each pixel with its motion profile: gives the probability distribution of the image velocity at each pixel in the image (used as a motion feature vector), successfully capturing both the direction and uncertainty:
$$\frac{1}{Z}\exp(-\alpha\text{SSD}[I^t(x_i), I^{t+1}(x_i + u)])$$
3. Normalised cut will give spatiotemporal volumes corresponding to different moving objects. Normalised cuts not only reflect similarity within a partition but also dissimilarity across partitions. We can take time slices of these volumes to indicate corresponding groups across time.
4. Steps:
    (a) The weight on the graph edge denotes the similarity in the pixels' motion profiles. The edge weight is given by cross-correlation between two motion profiles:
    $$\exp(-(1 - \sum_{dx} P_i(dx)P_j(dx))/\sigma_m^2)$$
        i. It should be noted that this measure of motion similarity will distinguish between two pixels which have exactly the same true motion, but where the brightness profiles are such that the associated motion uncertainties are very different. If one of the pixels is in a region of constant brightness and another in a region of rich texture this will happen. This is handled in a post-processing step.

    (b) The motion profile information between each pair of pixels is summarised and solved as an eigenvalue problem.

    (c) Using Normalised-cuts partition the graph with the eigenvector with the second smallest eigenvalue. Use partition only if it is stable (by checking the cut cost)
5. Since this paper groups pixels on based on the affinity of motion profile, a local measurement, it ignores global constraints and appears unstable in noisy sequences.

5 Paragios and Deriche [1999]

1. Main aim being simultaneous tracking of many non-rigid objects. Employs Geodesic Active Regions with a curve-based objective function having boundary and region-based terms.
2. Deals with motion estimation and tracking simultaneously.
3. The objective function is optimised (minimised) using gradient descent methods; the initially proposed curves are propagated toward the best partition under the influence of boundary and intensity and motion-based forces.
4. Boundary terms aims to find a minimal length contour attracted to region boundaries. Region-based terms in the objective function aims to maximise the quality of the segmentation map.

5. The motion detection module uses the difference frame to create parametric (Gaussian or Laplacian) distributions which can be used to model static (single component for static density) and mobile pixels (multi-component mobile density). The unknown parameters of the static and mobile component of the distribution are estimated using Maximum likelihood principle.

6. It assumes object motion can be described using global affine model $A(x, y)$ which is valid for a majority of object pixels.

7. The total Geodesic Active Region Tracking model is given by the sum of Boundary, motion detection, intensity and visual consistency module's objective function (see paper Section 3 for details on modules). Now given this function it will be minimised using a gradient descent method.

8. *Minimising with respect to the motion parameters* A_i only depends on visual consistency term. This is done iteratively - and each minimisation of an Euler-Lagrange equation helps updating the motion parameters $[A_i^{\tau+1} = A_i^{\tau} + \delta A_i]$.

9. *Minimising with respect to the Curve Position* involves optimising the boundary (has 2 terms: one that shrinks or expands the curve towards the object boundaries and another that attracts it to the object boundaries), motion detection (aims to shrink the curve when it is in a stationary background area and expands it when in a moving object), intensity segmentation (moves the curve in the direction which creates interior regions with desirable intensity properties), and visual consistency forces (moves curve in the direction which minimises the intensity error between the interior curve region and the object position in the previous frame) (all which act in the direction of the normal) in an Euler Lagrange system.

6 Ross [2000]

1. Thesis develops a technique to combine colour and motion features to produce optical flow and segmentation.

2.

3.

7 Stauffer and Grimson [2000]

1. How can motion tracking help to learn patterns of activity in the site.

2. Motion segmentation is done using an adaptive bg subtraction method where each pixel (process) is modelled as an adaptive mixture of Gaussians and uses an online approximations to update the model. These mixture of Gaussians are evaluates to see which ones come from a background process and which ones don't.

3. Aims to find a statistical interpretation of the activities in a scene (learning patterns in the scene).

4. The tracker should not only detect motion but also reliably return properties of an object like shape and size. The tracker should also be robust to a wide range of effects like lighting changes, moving trees, ans so on.

5. Each time the parameters are updated, they use a heuristic to find if the pixel belongs to a background process or not - pixel values that do not match any of the background Gaussians are grouped together using connected components. These connected components are tracked across time using a multiple hypothesis tracker.

6. Every new pixel X_t is checked against K Gaussian distributions until a match is found (match is within 2.5 std. dev. of a distribution). If none of the distributions match the current pixel, the least probable distribution is replaced by a new distribution with a mean of X_t with a high variance and low prior weight. These prior weights $w_{k,t}$ are adjusted using the learning parameter to make a causal low-pass filtered average of the posterior probability that pixel values have matched model k.

7. Gaussian distributions having more evidence (prior weight) and less variance (sort distributions by w/σ are taken as part of the background.
8. – left interpretation of motion tracks / classification of activities –

8 Yu and Shi [2001]

1. Exploits both region and boundary information for segmentation.
2. Boundary based methods for segmentation - Canny [5], Snakes [12], Balloons [6] - are bad at detecting faint contours. Region based groupings - region growing n merging [1] - since they lack a global cost function the boundaries might be noisy.
3. Uses two graphs - (1) pixel graph $G = (V, E)$ where weight on edges encode region cues (intensity, texture, motion similarity); (2) edgel graph $\partial G = (\partial V, \partial E)$ (the graph where nodes become edges) where weight on edges encode contour cues/boundary cues.
4. Consistent partitioning on the two graphs is guaranteed by the coupling of their partitioning membership vectors through edge-node incidence matrices. Such a link through linear operators allows us to derive analogous eigendecomposition solutions to the global optimisation criterion.
5. The main of partitioning is to maintain high within group affinity both for nodes in G and nodes in ∂G.
6. Use normalised-cuts on these G and ∂G graphs connected by an edge node matrix H which is essentially a difference operator (?)
7. Paper was expanded by Yu et al. [2002] to include a part-based recognition system detects object patches, supplies their partial segmentation as well as the spatial configuration between the segments. This patch grouping is done to find the set of patches that conforms best to the object configuration.

9 Paragios and Deriche [2002]

1. Many ideas same as 5
2. Tackles the frame partitioning problems in CV with N classes.
3. Like 5 it too exploits boundary based and region based techniques in a single framework. Their framework is called the **Geodesic Active Region** model "that incorporates boundary and region information sources under a curve-based minimisation framework to deal with frame partition problems."
4. This work expands on 5 level set formulation, by having a framework that works in **mutually exclusive propagating curves** - this is done by introducing artificial forces which penalises pixels which fall in multiple regions, or don't fall in a region at all.
5. Shows application in image segmentation, supervised texture segmentation and tracking.

10 Megret and DeMenthon [2002]

1. Three categories of spatio-temporal grouping techniques: (1) segmentation and spatial priority, (2) segmentation by trajectory grouping, (3) joint spatial and temporal segmentation.
2. Two main types of segmentation methods are used in videos, alone or in combination: motion-based methods, and colour/texture-based methods.
3. "Most approaches handle these spatial and temporal dimensions separately, making distinction between spatial segmentation, which groups features using spatial coherence criteria, and temporal tracking which groups features using temporal invariance hypothesis" - these leads to orders in which this can be done: segmentation with spatial priority [1-9]; and trajectory groupings [10-14].

4. A more recent method where both spatial and temporal segmentation are dealt in a unified framework - processes the spatio-temporal volume directly [15-18].

5. Features used could range from individual pixels, spatial regions resulting from colour/texture over-segmentation, geometric features such as interest points and edges.

6. Segmentation with Spatial priority:

    (a) Methods relying on motion: (1) Motion similarity methods where parameters are estimated locally; (2) Motion model fitting - involve the evaluation of asymmetric measures of the quality of fit of an element to the motion model.

    (b) Colour texture segmentation - useful in noisy data where motion fields might be unreliable and objects comprise of uniform colour/texture are and have high contrast. Here segmentations are matched temporally by seed propagation [7], and matching colour invariance and spatial overlap [29].

    (c) In these two techniques temporal coherence is enforced through: initialisation from previous frame [1,5,21,3,4,22] or explicit temporal constraints [30,21,31].

7. Trajectory grouping (used usually when objects undergo slow motion and/or have a lot of discriminating features):

    (a) Motion similarity: An interesting paper is [10] which forms spatio-temporal flow curves forming trajectories by their curvature and slope values in a fixed time interval. Features are clustered together through K-means on these trajectories. Analysis of merging and splitting of clusters gives info on occlusion events.

    (b) Grouping using explicit parametric models

8. Joint space and time (considers a spatio-temporal block):

    (a) Clustering in Feature space: [15] feature space: position (2), colour (3), time (1) - clustering through Gaussian mixture models in an EM framework; [16] pre-compute optical flow, feature space: motion angle (2), motion distance (2), colour (3) - clustering through mean-shift.

    (b) Graph based segmentation: graph formed from pixels as nodes and similarity measures as edges from spatio-temporal volume [41,42,6]. Also 4 comes in this category. Fowlkes et.al. [18] attaches a feature vector x_i to each pixel consisting of location (x,y,t), colour (L,a,b) and optical flow (u,v). The affinity between two pixels is computed as:

    $$W_{ij} = exp\big\{ -\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \sum^{-1} (\mathbf{x}_i - \mathbf{x}_j) \big\}$$

    where $\sum$ gives the weights of each feature. To reduce the complexity of the problem they use Nyström approximation of the normalised cut algorithm to compute the segmentation.

    [21],[3],[2],[4] Efficient spatio-temporal grouping using Nystrom method.

11 Wong and Spetsakis [2002]

1. Successively more elaborate models for optical flow and segmentation is done by extracting regions which are consistent in motion. The main idea is quite similar to 15

2. Essentially a motion segmentation method based on tracking.

3. The algorithm:

    (a) Give a small seed window (they give a 10×10 pixel window)in the first frame which falls inside the to-be-tracked object.

(b) Uses Least SSD to find the ideal optical flow vectors:

$$\text{LSSD}(u,v) = \min_{(u,v)} \sum_{x,y \in S} (I_{N-1}(y+u, x+v) - I_N(y,x))^2$$

where S is the seed window. The search window goes from $-30 \cdots 30$ pixels.

(c) Once we can align I_{N-1} to I_N, compute the sub-pixels optical flow by moving it successively in the 8 direction (by sub-pixels) and finding the location which gets the minimal SSD.

(d) Do segmentation on the above constant flow model to find which region to apply for affine flow computation.

(e) To find the affine flow, using a differential approach they compute the standard least squares for the following eq:

$$\text{SSD} = \sum_{\text{all } x,y} {}^{(N-1)}I_{\text{track}} \cdot \left[I_{N-1}(x,y) - I_N(x,y) + I_{N-1,x}(x,y) \begin{bmatrix} u_0 \\ u_x \\ u_y \end{bmatrix} + I_{N-1,y}(x,y) \begin{bmatrix} v_0 \\ v_x \\ v_y \end{bmatrix} \right]^2$$

where $u_0, u_x, u_y, v_0, v_x, v_y$ are the affine parameters, ${}^{(N-1)}I_{\text{track}}$ represents the region on which the affine optical flow is computed.

(f) By aligning all the previous images to the last image, they segment out the object to-be-tracked by thresholding the pixel-wise SSD. This is done because the SSD of the aligned tracked object would be small. This process can be made computationally tractable by aligning just the first and second moments of images (See paper for eq.s). The threshold is set by incorporating camera and motion noise.

(g) To remove erroneous regions a post-processing step is employed. It involves frame differencing to find moving pixels and then AND'ing them with the tracked frames (region-growing).

12 Felzenszwalb and Huttenlocher [2004]

1. Start with each pixel being a segment. From here we would merge segments/

2. The internal difference of a segment is defined as:

$$Int(C) = \max_{e=MST(C,E)w(e)}$$

where $MST(C,E)$ denotes the minimum spanning tree of the segment C with edges E each with a weight $w(e)$. The intuition they propose for this measure is that this MST can only remain connected if edges of atleast weight $Int(C)$ are considered.

3. The difference between two segments C_1, C_2 is defined as:

$$Dif(C_1, C_2) = \min_{v_i \in C_1, v_j \in C_2, (v_i, v_j) \in E} w(v_i, v_j)$$

i.e. the minimum weighted edge between two segments. If there is no edge $Dif(C_1, C_2) = \inf$

4. In each step we can merge segments if this criterion is met:

$$Dif(C_1, C_2) < \min(Int(C_1) + \tau(C_1), Int(C_2) + \tau(C_2)) \ = MInt(C_1, C_2)$$

where $\tau(C) = k/|C|$. This dictates if the difference between the different components is smaller than any of the two component edge differences, there is little chance of a boundary between the two components, hence they can be merged. The $\tau(C)$ is there to provide some support for merging only in absence of a strong edge - it also helps grow regions from single pixels ($|C|$ denotes the size of the component, and k is a constant).

5. Each step involves sorting un-connected edge strengths. Keep repeating until we can't make any more merges.

6. Discusses different edge weighting schemes $w(v_i, v_j)$, in which it chooses a function which is based on the absolute intensity difference between pixels connected by an edge. They show results using a feature space built on (x, y, r, g, b).

13 Zitnick et al. [2004]

1. IN THIS PAPER I HAVE CONCENTRATED MORE ON THE OVER-SEGMENTATION PART
2. Basically a view interpolation paper for dynamic scenes; for which it uses a colour segmentation technique as an input to the stereo algorithm to generate visually-consistent correspondences between different camera views.
3. Layered depth image representation.
4. There is no reference frame i.e. all images are treated equally and occlusions are modelled explicitly.
5. Not as concerned about the error in disparity as much as the error in intensity.
6. Some works in stereo use segmentation: Tao et al. [2001] uses a planar constraint; Zhang and Kambhamettu [2001] uses segments for local support.
7. Disparity within segments must be smooth but need not be planar.
8. Segmentation
   (a) Supposes that disparity discontinuities coincide with colour discontinuities.
   (b) Smooth image by a variant of anisotropic diffusion. Each pixel is its own segment at start.
   (c) Merge 4-connected neighbour segments into a single segment if the euclidean distance of their average colour values is less than 6.
   (d) Segments smaller than 100 pixels are merged with a neighbouring segment with the closest average colour.
   (e) Segments with homogenous colour larger than 40 pixels in length, we split them up.
9. Initially a single disparity label is assigned to each segment. Later the disparities are smoothed out by re-projecting pixels to neighbouring images and taking the average disparity across these frames.

14 Xiao and Shah [2005]

1. Assumes planar regions and extracts a set of affine or projective transformations that these regions undergo, detects the occlusion pixels and segments scene into motion layers
2. Argues that small patches don't give reliable motion parameters due to the over fitting problem and therefore employs a step before clustering regions into layers.
3. These are the major steps in the algorithm:
   (a) Detect seed correspondences in 3-5 frames and expand these initial square seed regions to arbitrary shaped regions and reject outliers using graph-cuts integrated with level set representation.
      i. Use Haris corners and KLT to track corners over a short period in a 17×17 window.
      ii. Gradual expansion of seed regions by identifying the supporting pixels by bi partitioning graph cuts method with level set representation - having a smoothness energy term:

$$\begin{aligned} E &= E_{\text{smooth}}(f) + E_{\text{data}}(f) \\ &= \sum_{(p,q)\in\mathcal{N}} V(p,q).T(f_p \neq f_q) + \sum_{p\in\mathcal{P}} D(p, f_p) \end{aligned}$$

where E_{smooth} is the piecewise smoothness term, E_{data} is the data error term, $\mathcal{P}$ is the set of pixels in the image, $V(p,q)$ is a smoothness penalty function, $D(p, f_p)$ is a data penalty function, $\mathcal{N}$ is a four-neighbour system, f_p is the label of a pixel p, and $T(.)$ is 1 if its argument is true and 0 otherwise. In this bi partitioning problem, the label f_p is either 0 or 1.

iii. The level set representation evolves the contour in the normal ⊥ direction of the seed region. Use the contour of the seed region as a prior to compute the level set, v, of this region. Then, apply v to change the weight of the links on the sink side. Therefore, we effectively restrict the graph cuts algorithm to gradually expand the seed region .

(b) In two steps, we merge these regions into layers on basis of motion similarity. Here an occlusion order constraint is imposed to find the occlusion regions.

i. Merge two regions R_1, R_2 of the initial layers if the SSD of the two regions after applying the transformation H_2 on R_1 results in a majority of the pixels in R_1 supporting R_2. The motion parameters are recomputed after the merger.

ii. After that, the bi partitioning graph cuts algorithm again used to prune the un-supporting pixels from the new region.

iii. The result is each pixel belonging to a layer label showing some coherent motion.

(c) The correct layer segmentation is found by using a graph cuts algorithm and occlusions between overlapping layers are explicitly determined i.e. maintain consistency of layer segmentation using the occlusion order constraints. All of this is done over multiple frames.

i. – Three state pixel graph –

ii. – Multiframe motion segmentation via Graph Cuts –

iii. – Energy minimisation with occlusion order constraints –

4. The output is in form of a segmentation in terms of layers and their 2D motion parameters.

15 Galun et al. [2005]

1. Starts off with local ambiguous optical flow measure. The ambiguities are resolved by aggregation to get reliable motion estimates (iteratively employing more complex motion models - first translation, then affine, and finally recovering 3D motion (described by fundamental matrix)).

2. The method is also integrated with segmentation method using intensity cues.

3. Aims to combine motion with intensity cues, mainly for small motions.

4. Fine to coarse aggregates:

    (a) Clustering: Select seed elements; assign elements from previous levels

    i. Optical flow initially computed as 4. Their method to comparing motion profile is similar to 4 except that they smooth the profiles before comparing.

    ii. Define a strength of association for fine element to a coarse seed

    (b) Re-estimation: estimate the common motion of the cluster (the motion model is decided by the "level of scale") (For translational motion we find raw motion cues - getting the translational motion of the centre of the mass of the aggregate)

    i. Each coarsening step begins by selecting a subset of the elements from the previous level as seeds, with the constraint that all other elements are strongly associated with (subsets of) these seeds. To aggregate the motion profile is computed by multiplying all the child motion profiles (see exact function!) - this technique gives sharply peaked motion profiles in textured regions in just 1-2 coarsening steps - but it doesn't work in uniform regions.

    ii. Check peaking of seeds and accumulate moments of those seeds

    iii. If there is enough statistics, calculate affine transformation by merging moments from peaked and bar-peaked profiles.

    iv. If there is enough statistics, calculate fundamental matrix from peaked profiles

5. Combine motion with intensity cues by using the above motion segmentation algorithm with Segmentation by Weighted Aggregation (SWA) [13].

    (a) SWA is multiscale graph partitioning algorithm (saliency measure?).

16 Ogale et al. [2005]

1. Uses occlusion to figure motion segmentation in a scene where there is motion due to camera and due to the scene elements. Aims to find depth ordering using occlusions.

2. Given a noisy flow field, any motion estimation technique will yield a region of solutions in the space of translations instead of a unique solution; they refer to this region as the **motion valley**.

3. Types of object motions:

   (a) The background objects motion is different from the motion of objects of interest. Here motion-based clustering will work.

      i. Perform phase correlation to find translational, scaling and rotational parameters - this will extract background motion - set select some points S which represents it.

      ii. Optical flow values of S help estimate background *motion valley* using the 3D motion estimation technique in [25]. Reprojecting the flow back to the image plane, we compare the flow to actual flow, and Class 1 objects are wherever there is large disparity in the flows.

   (b) Here the background motion and object of interest move similarly. Here we will need to use an ordinal depth conflict constraint to resolve the motion - in which we will need to resolve the ordinal depth of the object.

   (c) In this class we can't take help from finding ordinal depth - we will need to do cardinal comparisons from structure from motion and structure from X (paper uses structure from stereo). Using k-means clustering on depth ratios ($k = 3$), it detects the background - the largest cluster. Pixels with depth ratios greater or less than the background are the objects we are trying to find.

4. Since optical flow estimation provides us with a segmentation of the scene (regions of continuous flow), we now have to assign flows to the occluded regions and merge them with existing segments to find ordinal depth.

5. Using 3 frames F_1, F_2, F_3 we can find the optical flow u_{12} and u_{23} and the reverse optical flow u_{21} and u_{32}. In this process we also know regions of occlusion O_{12} (regions present in F_1 but not in F_2), and O_{23}. The aim is to find the ordinal depth of the regions in the frame. Consider finding the region labels of occluded regions in F_3, O_{23}. We know that regions occluded in F_3 would have been visible in F_1, F_2; hence we can use segmentation of flow of u_{21} to find the labels for regions that subsequently got hidden in F_3 i.e. O_{23}. Now finding ordinal depth is trivial.

17 Zitnick et al. [2005]

1. Optical flow and segmentation while accounting for matting in overlapping regions (modelled with α). Simultaneous estimation of flow, segmentation and matting in a single generative model depending on appearance and motion constraints.

2. Bidirectional motion estimated through spatial coherence and similarity of segment colours.

3. Temporally consistent segmentation - it argues that segmentation from single image just using colour information can be ambiguous. Temporal consistency involves ensuring similarity of segment shape and colour across time as well as spatial coherence.

4. The paper shows correct motion segmentation's application in frame interpolation.

5. Proposes fine grade segmentation with translational motion model. Proposes a matting model similar to 21 where each pixel can belong to two segments (out of K segments) with a corresponding association of α and $1 - \alpha$ (s_i^1, s_i^2 denotes the two segments pixel i belongs to - $c_i \approx \alpha_i c_{i,s_i^1} + (1 - \alpha_i) c_{i,s_i^2}$).

6. The parametrisation of the variability in the data corresponds to a generative model of a single pixel, which generates pixel colours and positions by the following hierarchical statistical process. First hidden segment pairs are sampled from a uniform distribution. Then two hidden pixel colours are generated c_{i,s_i^1}, c_{i,s_i^2}

by observing the position r_i of the pixel. Then the α_i value is generated from a prior distribution which favours values close to one. The last step in the generative process is generating the observed pixel colour c_i by noisy alpha blending of the two parent segment colours.

7. Colour and coordinate variation of a segment k decided by the parameters $\phi_k = (\mu_k, \sum_k, \eta_k, \Delta_k)$, where first two parameters define the Gaussian model for the colour, and the Gaussian model for the spatial distribution of the segment's pixels is defined by the last two.
8. For segment correspondence (aiming for consistent segmentation) two mappings are defined; M^{tu} (mapping from image X^t to X^u) and M^{ut}. To cater for occlusions and dis-occlusion events $m_k^{tu} = j$ doesn't always imply $m_j^{ut} = k$. Although each segment should have some mapping.
9. Segmentation consistency and displacement variables are computed exactly the way they were used in 21

18 Zelnik-Manor et al. [2006]

1. Tracking by subspace methods applied directly to features pixel intensities; assumes orthographic projection and their generalisation to perspective projection.
2. Motion segmentation through temporal consistency (in terms of multi-frame linear subspace constraints) of behaviour across multiple frames.
3. Applies multi-body segmentation to a flow field matrix $[U, V]$ (where both U and V are directional motion vectors of Frames × Pixels dimensions) giving rise to multi-body factorisation with directional uncertainty.
4. In segmentation of $[U, V]$, the decision of grouping together two objects is based on the ranks of the matrix - which exploits the linear dependency of flow fields of a single object. Since the flow-field matrix $[U, V]$ is of rank 1, there exists a set of basis trajectory vectors and a set of basis flow-fields such that they can be factored into two matrices of flow coefficients and flow basis.
5. Since tracking might not be reliable for all feature points, directional uncertainty is introduced by a weight matrix $[U, V]Q$
6. Segmenting the entire image into objects can be done by sorting the columns of the brightness-measurement matrix $[G, H]$ which is done by finding its RREF. Since the rank of $[G, H]$ is considered to be small, RREF is computed on the SVD of $[G, H]$.

19 Zitnick and Kang [2007]

1. IN THIS PAPER I HAVE CONCENTRATED MORE ON THE OVER-SEGMENTATION PART
2. Stereo method designed for image based rendering - where the stereo algorithm depends on over-segmenting images.

   Computes match values for entire segments rather than individual pixels providing robustness to noise and intensity bias. Colour-based segmentation also helps reducing boundary artifacts and works well in texture-less regions. The side-effect being that depth discontinuities occur at colour boundaries.
3. The depth of each segment is computed using loopy belief propagation in an MRF framework.
4. The method is in contrast to other works since its measure of success isn't minimising the disparity with the actual depth but to make interpolated views using the extracted depth distributions look physically correct.
5. "More recent work directly relies on colour segmentation (Tao et al., 2001), and over-segmented regions (Zitnick et al., 2004)."
6. Discusses the entire spectrum of segmentation: starting from small pixel-wise models like SSD, over-segmentation - to global motion models which take the whole image as a segment (lets say in panoramic images)

7. “If a segment is too small, it is difficult for it to un-ambiguously ?nd the correct pixel correspondence. As a result, some mechanism for using information from neighbouring segments is typically required to reduce the ambiguity. For single pixel correspondence, graph cuts (Boykov et al., 2001) or belief propagation (Sun et al., 2003) provides this mechanism. In Tao et al. (2001), neighbouring segments are used to help in finding correct correspondences in the vicinity of occlusions.”

8. it uses each segment as a node in the MRF graph and disparity levels as states. The disparity space image values at nodes are updated using loopy belief propagation. It eventually iteratively shares disparity information across images while updating the segments’ disparity beliefs. When the solution converges, each segment is assigned to the disparity with the maximum probability belief.

   (a) The aim of the over-segmentation is to split the image into regions likely to have the similar disparities. It supposes disparity discontinuities usually coincide with intensity edges.

   (b) smoothing as done in 13

   (c) the image is partitioned into 8 × 8 pixel blocks

   (d) K-means is run using these parameters: mean of colour space; spatial extent of the segment with Gaussian mean and variance.

   (e) A segment smaller than 10 pixels is discarded.

20 Kumar et al. [2008]

1. Unsupervised approach to generative layered representation for motion segmentation.

2. Explicit modelling of spatial continuity to give contiguous segments. It also models occlusions; it handles multiple frames; and the model is learnt from independent keyframes.

3. Since a layered representation is used, any frame in the sequence can generated from the learnt model by assigning appropriate values to its parameters and latent variables. This model is called the **latent image** which essentially holds all the parameters of the model.

4. Latent image: n_p segments; each p_i segment has a binary matte Θ_{Mi}, an appearance parameter given by its RGB values at pixel x represented as $\Theta_{Ai}(x)$ (the colour histogram for the segment is $\mathcal{H}_i$), each segment p_i is given a layer number l_i - and p_i can occlude p_k only if $l_i > l_k$ - in short the latent image is defined by mattes Θ_M, appearance Θ_A, its histograms $\mathcal{H}$, and the layer numbers l_i for the n_p segments.

5. Using the generative model we can generate any frame using the segment’s latent variables Θ_{Ti} transformation parameter (x,y-translation, x,y-scaling, and rotation), and the lighting parameters $\Theta_{Li} = a_i, b_i$. Hence a complete representation is done by $\Theta = \{n_p, \Theta_M, \Theta_A, \mathcal{H}_i, l_i; \Theta_T, \Theta_L\}$

6. Given a video, the best layered representation is the one which minimizes the energy of the layered representation, which has an appearance term, prior term (encouraging spatial continuity), and data-dependent contrast term (encouraging boundaries of segments to lie on image edges).

7. The initial estimate of the model parameters is done by dividing the frame into small rectangular patches and determining their transformations. These patches are put in an MRF framework which estimates segment motion with discontinuities using a loopy belief propagation technique (a prior is defined with a simple Potts model which penalises neighbouring patches not moving rigidly).

8. Given this initial estimate, the shape of the segments along with the layering is learnt by minimising the objective function using $\alpha\beta$-swap and α expansion techniques. This step is done segment at a time. $\alpha\beta$ swap involves swapping pixels in the nearby segments to see which is the right configuration. In α expansion we expand each segment and check overlaps in segments to see if either A occluding B reduced the energy or the other way round.

9. Now the appearance model of each segment is updated; the transformed model is updated by searching in the neighbourhood and choosing a transformation which minimises the SSD.

21 Zitnick et al. [2009]

1. Algo. steps:

    (a) Compute segmentation for all image $F_{1\cdots n}$. The segments are relatively small to previous techniques - hence the optical flow of a segment can be simply represented by a translation.

    Quadtree approach; Divide image into 60×60 pixel blocks; keep breaking each block into 4 equal blocks until either the colour variation in that block is below a particular threshold or the size of the block is small enough.

    (b) Compute optical flow in both directions for all consecutive image pairs $F_n F_{n+1}$ and $F_{n+1} F_n$.

    The initial optical flow of all segments is set as zero vectors

    (c) Loop over the following steps until prescribed iterations done:

    i. Refine the segmentations using the *refined* optical flow estimate from the last step.

    A. Concept: A pixel's **main segment** is one to which it gives it most colour contribution; whereas a **secondary segment** is an adjacent segment to which the pixel contributes lesser colour. Border pixels have an α value associated with the main segment, and $1 - \alpha$ to the secondary segment. Pixels in the interior of the segment have $\alpha \approx 1$. $C_{\text{pixel}} = \alpha C_{\text{main}} + (1 - \alpha) C_{\text{secondary}}$. $C_{\text{main}}, C_{\text{secondary}}$ are average colours of a segment.

    B. In each refinement stage each pixel is selected and we try to guess what could be its main segment and secondary segment. Both these segments should lie in lets say a 5×5 pixel neighbourhood. If there is only one segment in that neighbourhood, it is selected as the main segment and the pixel is given an $\alpha = 1$. If more than one segment exist we need to go through a scoring mechanism: (1) taking each pair of segments we see how many pixels in that neighbourhood have set either of the two segments as the main segment divided number of pixels in the proposed segment, (2) using the flow of the current main segment of the pixel, it is projected to the next image - and for each segment pair in a 3×3 pixel neighbourhood the same tally is computed divided number of pixels in the proposed segment. (3) We compute an alpha similarity score as: $e^{-R(p,S_{i,k},S_{i,l})^2/\sigma_s^2}$, where R is the perpendicular distance from the line passing through the average colours $S_{i,k}, S_{i,l}$ of the two segments. σ_s is the standard deviation of the variation in colour of the main segment. If the perpendicular residual line is not between the line connecting the two segments, then the main segment is the one closest to the pixel colour point, and there is no secondary segment.

    C. Multiply the 3 scores to get the total score. The segment pair with the highest score is selected as the winning pair. The pixel is assigned to the main segment. The new α value is computed as given in 1(c)iA.

    ii. Refine optical flow using the set of segmentations

    A. For most segments the refined flow is taken as the distance between the centroids of the two segments.

    B. The set of corresponding segments is found in the two images by the following score:

    $$e^{-(\Delta C)^2/\sigma_c^2} T(S_{i,k}, S_{j,l}) e^{((\Delta x) - \bar{v}(S_{i,k}))^2/\sigma_x^2}$$

    ΔC difference in average colour of two segments, σ_c is the estimated standard deviation of the difference of average colours in segments, $T(S_{i,k}, S_{j,l})$ is the ratio of the number of pixels in the two segments (dividing by the bigger segment), Δx difference in the position of centroids in the two segments, $\bar{v}(S_{i,k})$ is the weighted average flow, and σ_x is the estimated standard deviation in the difference of the position of centroids.

    C. The segment having the highest score is chosen as the corresponding segment.

    iii. This process can be reversed too - refine segmentation first then refine optical flow.

2. – Overlapping Segments as a Generative Model of a single image –

(1) Wang and Adelson propose the use of optical flow to estimate the motion layers, where each layer corresponds to a smooth motion field [27]. (2) Wang and Adelson [18] were arguably the first to develop this idea, usingthe affine model for flow. They proposed an iterative approach to create and remove segments, based on the pixel-wise flow computed using a method similar to [2]

Ayer and Sawhney combine Minimum Description Length (MDL) and Maximum-Likelihood Estimation (MLE) in Expectation-Maximization (EM) framework to estimate the number of layers and the motion model parameters for each layer [1].

References

MJ Black and AD Jepson. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):972–986, 1996.

M.M. Chang, A.M. Tekalp, and M.I. Sezan. Simultaneous motion estimation and segmentation. *IEEE Transactions on Image Processing*, 6(9):1326 –1333, sep 1997.

P.F. Felzenszwalb and D.P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2):167–181, 2004.

M. Galun, A. Apartsin, and R. Basri. Multiscale segmentation by combining motion and intensity cues. In *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '05)*, volume 1, pages 256–263, june 2005.

F. Heitz and P. Bouthemy. Multimodal estimation of discontinuous optical flow using Markovrandom fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(12):1217–1232, 1993.

M.P. Kumar, P.H.S. Torr, and A. Zisserman. Learning layered motion segmentations of video. *International Journal of Computer Vision*, 76(3):301–319, 2008.

R. Megret and D. DeMenthon. A Survey of Spatio-Temporal Grouping Techniques. Technical Report CS-TR-4403, Language and Media Processing, University of Maryland, College Park, MD, August 2002.

A.S. Ogale, C. Fermuller, and Y. Aloimonos. Motion segmentation using occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):988–992, 2005.

N. Paragios and R. Deriche. Geodesic active regions for motion estimation and tracking. In *Proceedings of the Seventh IEEE International Conference on Computer Vision, (ICCV '99)*, volume 1, pages 688–694, 1999.

N. Paragios and R. Deriche. Geodesic Active Regions: A New Framework to Deal with Frame Partition Problems in Computer Vision. *Journal of Visual Communication and Image Representation*, 13(1-2):249–268, 2002.

M.G. Ross. Exploiting Texture-Motion Duality in Optical Flow and Image Segmentation. Master's thesis, Massachusetts Institute of Technology, Aug 2000.

J. Shi and J. Malik. Motion segmentation and tracking using normalized cuts. In *Proceedings of the Sixth IEEE International Conference on Computer Vision, (ICCV '98)*, pages 1154 –1160, jan 1998.

C. Stauffer and W.E.L. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

K.Y. Wong and M.E. Spetsakis. Motion segmentation and tracking. In *Proceedings of 15th International Conference on Vision Interface*, pages 80–87, 2002.

J. Xiao and M. Shah. Motion layer extraction in the presence of occlusion using graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10):1644–1659, 2005.

S.X. Yu and J. Shi. Perceiving Shapes through Region and Boundary Interaction. Technical Report CMU-RI-TR-01-21, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, July 2001.

S.X. Yu, R. Gross, and J. Shi. Concurrent Object Recognition and Segmentation by Graph Partitioning. In *Advances in neural information processing systems 15*, pages 1383–1390. MIT Press, 2002.

L. Zelnik-Manor, M. Machline, and M. Irani. Multi-body factorization with uncertainty: Revisiting motion consistency. *International Journal of Computer Vision*, 68(1):27–41, 2006.

C.L. Zitnick and S.B. Kang. Stereo for image-based rendering using image over-segmentation. *International Journal of Computer Vision*, 75(1):49–65, 2007.

C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 23(3):600–608, 2004.

C.L. Zitnick, S.B. Kang, and N. Jojic. Simultaneous optical flow estimation and image segmentation. US Patent 7,522,749, Apr 2009. Filed: July 2005.

C.W. Zitnick, N. Jojic, and Sing Bing Kang. Consistent segmentation for optical flow estimation. In *Proceedings of the Tenth IEEE International Conference on Computer Vision, (ICCV '05)*, volume 2, pages 1308–1315, oct 2005.