Learning to Perceive Coherent Objects

Nimrod Dorfman[1] (nimrodd@weizmann.ac.il)
Daniel Harari[1] (danny.harari@weizmann.ac.il)
Shimon Ullman[1] (shimon.ullman@weizmann.ac.il)
[1]Department of Computer Science, Weizmann Institute of Science, Rehovot, Israel

Abstract

Object segregation in a visual scene is a complex perceptual process that relies on the integration of multiple cues. The task is computationally challenging, and even the best performing models fall significantly short of human performance. Infants initially have a surprisingly impoverished set of segregation cues and their ability to perform object segregation in static images is severely limited. Major questions that arise are therefore how the rich set of useful cues is learned, and what initial capacities make this learning possible. Here we present a computational model that initially incorporates only two basic capacities known to exist at an early age: the grouping of image regions by common motion and the detection of motion discontinuities. The model then learns significant aspects of object segregation in static images in an entirely unsupervised manner by observing videos of objects in motion. Implications of the model to infant learning and to the future development of object segregation models are discussed.

Keywords: Visual perception; computational modeling; development; object segregation; figure-ground.

Background and Goals

We naturally perceive the scene around us as containing coherent objects, separated from each other and from their background. Even in a complex image such as Figure 1A, we can count for example the number of distinct cars, delineate their boundaries, etc. The ability to segregate the scene into objects, delineate their boundaries, and determine occlusion relations (termed here 'object segregation'), relies on a complex set of processes, which integrate multiple cues that are only partially understood.

Infants' initial ability to segregate scenes into coherent objects is rudimentary and it does not make use of even basic salient 'Gestalt' properties such as uniformity of texture, brightness or color, the smooth continuity of boundary contours, occlusion cues and the like (Spelke *et al*. 1993). For instance, infants at 3 months of age do not appear to distinguish that the shape in Figure 1B is likely to be composed of two distinct components. The contrast between Figures 1A and 1B illustrates the span of learning accomplished in performing object segregation. The ability to segregate objects based on multiple cues develops quickly already in the first year of life, but the learning process continues over an extended period of time (Kovaks *et al*. 1999). The process of learning object segregation raises fundamental questions for cognitive development and computational modeling of vision. For cognitive development, it is of basic interest to understand the innate capacities and learning mechanisms that allow the system to start from a surprisingly limited capacity for segregating the world into coherent objects, and reach the capability of the adult system. For computational modeling of vision, an intriguing possibility is to try to surpass the capabilities of current models by following a strategy similar to human development, namely, start with the appropriate set of basic capacities and learning mechanisms and allow the model to develop on its own the final segregation capabilities.

A B

Figure 1: Object segregation, infant to adult capacity. (A): A complex scene, easily segregated by an adult. (B): At 3 months, infants do not appear to divide the figure into two components (after Spelke *et al.* 1993).

In the current study we focus on specific sub-problems within this broad domain. We develop a model that incorporates simple basic capacities, which are known empirically to already exist in young infants. It uses them to segregate familiar objects and to extract and use so-called 'boundary ownership' cues (indicating boundaries as well as figure/background direction) for static object segregation. The model initially has no ability to segregate objects in static images, but it can compute visual motion and motion discontinuities. It is exposed in an unsupervised manner to video sequences containing moving objects. It uses them to segregate familiar objects in static images and to learn local boundary ownership cues. These are used as cues for static object segregation, applied to novel objects.

In the next sections we briefly summarize relevant background from developmental studies of object segregation, followed by a presentation of the current model.

Early Development of Object Segregation

Initial object segregation by infants is based almost exclusively on dynamic cues, which are then used to learn static object segregation. We focus below on two main

aspects of using visual motion for object segregation: grouping by common motion, and the use of motion discontinuities. We also comment briefly on the use of static cues.

Common Motion Infants use visual motion to group together adjacent regions that move together. These grouped image entities, discovered through motion, are also stored in memory and can subsequently be identified in static images (Needham 2001, Needham & Baillargeon 1998, Needham & Modi 1999, Spelke 1990, Spelke *et al*. 1989). For example, if 4.5 months old see in a static image a region A next to a second region B, their expectations are shaped by their recent experience of seeing these regions in motion. If A and B moved together, infants will treat them as a unit and will be surprised if they move separately, but not if they saw A or B moving alone. The grouping of regions into a single unit depends on their common motion: if two regions differ in their image motion, even if they remain in contact, they are treated as separate objects (Spelke 1990, Spelke *et al*. 1989). Retention in memory of the formed unit is limited in time (about 24 hours at 4.5 months of age), but grows gradually with age (Needham & Baillargeon 1998, Needham & Modi 1999). This use of stored object representations for segregation is termed 'object-based segregation', and it can generalize with more experience to other similar objects ('class based' segmentation), provided that the differences are initially small (Needham & Modi, 1999). Two regions moving together can also be grouped together to form a single unit when they are non-contiguous but separated behind an occluder (Kellman & Spelke 1983) provided that the parts are roughly aligned (Johnson & Aslin 1996).

Motion Discontinuities In addition to region grouping based on common motion, infants are also sensitive from an early age (5 months or earlier) to dynamic cues created by the boundaries of moving objects (e.g., Granrud *et al*. 1984).

Static Cues In terms of static cues, at 3-5 months contiguous regions that are not separated by a visible gap tend to be grouped together, and are expected for example to move together rather than separately (Needham & Baillargeon 1998, Spelke 1990, Spelke *et al*. 1989). At this age they show little or no evidence for using grouping principles based on uniformity of color, texture, and continuity of bounding contour in object perception. At 9 months the effect of such grouping cues is still weak (Spelke *et al*. 1993). The learning of static cues is gradual, and appears to depend on familiarity with many objects (Needham & Modi 1989, Spelke 1990).

Following extended learning, perceptual organization into distinct objects and their boundaries develops into a complex process that relies on a rich set of cues. In addition to image-based, or bottom-up properties, organization into objects depends on top-down cues, based on familiarity with specific objects and object classes. The different cues and their integration into a full segregation scheme are still a subject of active research in both human studies and computational modeling. Yearly competitions and evaluations of natural image segmentation[1] show consistent improvements, but current performance is still significantly below human performance. Due to space limitations, we will not review here different modeling efforts. The closest to the current study is the SANE (segmentation according to natural examples) model by Ross *et al*. (2009), where, like in the current study, motion segmentation was used to guide static segmentation. However, the SANE model does not use the two main components of the current model: learning boundary-ownership cues near a boundary, and learning object-based segregation. It uses instead local binary 5×5 boundary elements, with no ownership information, and their pair-wise relationships.

Goals of the Current Study

As reviewed above, infants are sensitive to motion cues for segregation, but lack sensitivity to most static cues for objects identity. It is therefore natural to ask how static segregation cues may be learned during development, guided by dynamic cues. We focus on two dynamic cues that are prominent in early infant perception. The first is common motion, guiding object-based segmentation. That is, infants naturally segregate adjacent image regions that share common motion, and can identify similar configurations in static images. One goal is therefore to model this learning of object-based segregation. Second, infants are sensitive to dynamic cues created by the boundaries of moving objects, and these are used by the model to learn useful static boundary cues. Although boundary ownership cues appear to play a major role in human object segregation (e.g. McDermott 2004, Ghose & Palmer, 2010), they are not usually used in computational models, in part because it is still unclear which features are useful for assigning boundary ownership. A possible outcome of a model for the unsupervised learning of boundary ownership features could be, therefore, the extraction and use of such features in future segmentation models and algorithms.

The Model

The current learning model has initially two 'innate' capacities for using visual motion to learn object segregation. The first is the capability to group together adjacent regions based on their common motion. A representation of the grouped shape is stored and can then be used for segregating similar shapes in novel static images. The second capacity is to extract motion discontinuities. These are used as teaching signals to extract image features located along object boundaries, together with a labeling of the figure/background sides, and

[1] http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

subsequently use them to locate novel object boundaries and identify the figure direction in new static images. These two components and how they are used by the learning model are described in subsequent sections, following a brief description of the training data used for learning.

Training and Testing Data

Data consisted of 48 movies, each depicting an object (doll, banana, remote control etc.) moved by hand in front of a textured background (12 objects, 12 backgrounds). For each movie, there are 3 other movies showing the same object on a different background, 3 movies showing different objects on the same background; the remaining 41 have different object and a background. Each movie is one minute long (1500 frames), frame size varies between 520×720 pixels to 576×752 pixels.

Object-based Segregation

The goal of object-based segregation is to learn the appearance of a specific object, such as the doll, fruit, etc., in our movies, and then find the full extent of the object and separate it from its background under new settings. The part of the model that deals with object-based segregation is based on an object detection model used, with some variations, in computer vision schemes, termed 'star model'. For the purpose of object segregation, the model is augmented with a 'back projection' stage. Since this part relies on existing object detection models it will be described here briefly.

The input to the object-based segregation is an image in a movie, together with the visual motion associated with the image. The scheme used for motion computation was an available optical flow algorithm (Sun *et al*. 2010) combined with background subtraction, assuming that the camera itself is stationary (as in Ross & Kaelbling, 2009).

The motion computation divides the image into two components: a stationary one, and a set of one or more moving regions. One of the moving regions is selected for further processing. The selected region is covered by local image descriptors, each one representing the appearance of a local region. The implementation used the standard SIFT image descriptor (Lowe 2004) because computationally, it is robust and efficient, and biologically, it is similar to intermediate level units used in modeling (e.g. S2 units in the cortical H-Max model, Riesenhuber & Poggio 1999). A single reference point C is selected at the center of the selected region, and for each image descriptor F_i, the displacement V_i from its location to the center C is stored. The object defined by the moving region is therefore represented by its center C, and the set of image descriptors (F_i), each one with its displacement V_i from the object's center.

Figure 2: Examples of object-based segregations produced by the algorithm. Bottom right: an erroneous example.

Segregation of Static Images If the same or similar object appears in a new image, it can be detected and segregated based on the above representation, using the following algorithm. The new image is represented by its local SIFT descriptors. For each descriptor F in the image, we find its K=25 nearest neighbors among the descriptors of the stored object. Each neighbor F_k, votes for the location of the center C according to the displacement V_k. Votes are weighted by the similarity between F and F_k, and aggregated over the image. If an image location C obtains a sufficient number of total votes, an object is detected, centered at location C. The full object is then segregated by a 'back projection' step: all image descriptors that contributed their votes to the selected location are identified as components of the detected object. A final object/background decision is made by an automatically set threshold.

Results – Speed and Generalization In infants, even a few seconds of observing an object in motion already affects subsequent segregation of the same object in static images (e.g. Needham, & Baillargeon 1998). The segmentation is effective for images of the same or similar object and generalizes gradually to less similar objects (Needham, & Baillargeon 1998, Needham & Modi 1999). Object-based segregation in the model showed similar characteristics. Brief (5 seconds) training was sufficient for learning object segregation of a specific object in subsequent parts of the movie, with some generalization to a different pose and different background. The object is often grouped by motion with the holding hand; the two can be separated when the hand is learned as an object on its own (Ullman *et al.* 2012). Figure 2 shows example segregations.

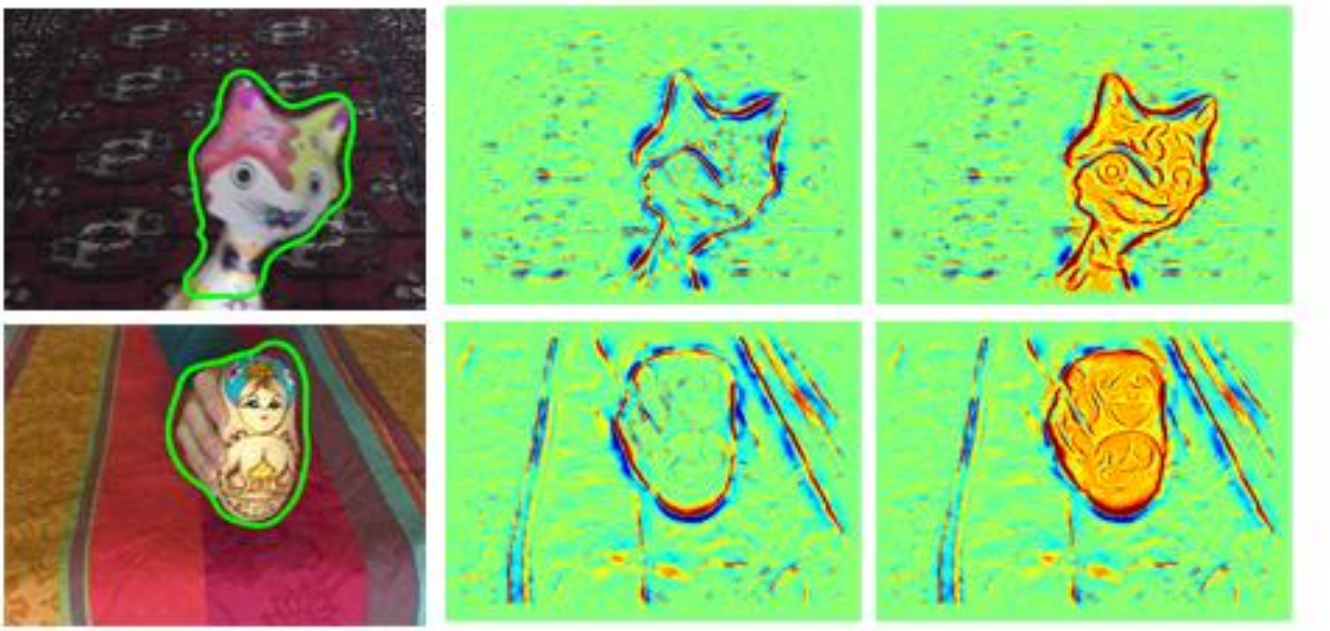

Figure 3: Detecting object boundaries. Left: Original image, with object-based segregation. Object is located, but boundaries are inaccurate. Center: Detection of boundary features. Warm colors indicate figure side of boundary, cold colors – ground side. Both object and background were not seen during training of boundary detector. Right: Combining object-based segregation with boundary detections. Object is detected with correct boundaries.

Results were tested by learning an object model in each movie using 5 sec and 40 sec segments, and testing on both later parts of the same movie, as well as the same object in other movies, with different backgrounds and larger variations in pose and lighting. Agreement between the true object (extracted by motion) and the model segregation were measured by the standard score $s = |T \cap S|/|T \cup S|$, where T is the true object and S the segmented. Mean scores for 5-sec training were s = 0.3 vs. 0.23 on same vs. different movies, and for 40-sec training s = 0.49, 0.36 respectively. Effects of training time and generalization are highly significant (1-tailed t-test, n=1200, $p < 10^{-6}$ in all comparisons).

The object-based segregation in the model segregates the general object region but it does not accurately delineate the boundaries. Since the object is represented by local appearance patches, it is sensitive to texture properties inside the object, in agreement with infant's object-based segregation (Needham & Modi, 1999). In contrast, the model shows limited accuracy around object boundaries; it will be interesting to test this prediction in infants' vision (see discussion).

Learning Boundary Features

The accurate delineation of boundaries is important for interacting with objects, e.g. for grabbing, finding free space to place them, etc. This is obtained in the model by a second mechanism, which uses motion discontinuities to learn static cues for occluding boundaries, as described next.

Learning Process To learn useful boundary features, motion discontinuities are used to guide the extraction of static boundary features and their figure-ground labeling. The learning procedure is simple, proceeding along the following stages. In each frame of the training movies, motion discontinuities are detected, and at each pixel along the boundary, image patches are extracted at 5 different sizes (ranging from 12×12 pixels to 60×60 pixels). Each patch is represented by a rotation invariant SIFT descriptor, producing a fixed-size descriptor regardless of original patch size. The motion signal is also used to label the figure part (which is moving in the training images) and background part (which is stationary) in each stored patch. From these, a subset of boundary patches is later selected, as described in the Results section below.

Use In Static Images The learned boundary features are then used to identify likely object boundaries in novel static images. Given a static input image, local SIFT features are extracted at the same 5 sizes, densely over the entire image. For each feature, its 25 nearest neighbors in the stored set of trained boundary features are extracted (using a fast approximation algorithm, Arya & Mount 1993). These neighbors are used to estimate the likelihood of an object boundary at this location, and to identify the figure side of the potential boundary. Specifically, each neighbor i has a SIFT descriptor D_i and an object direction θ_i. For an image patch with descriptor D, we define the predicted object direction θ and a score S as follows:

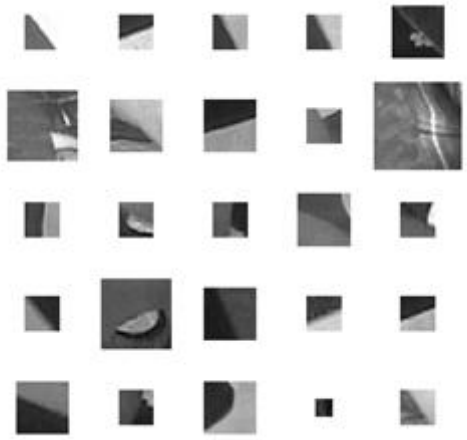

Figure 4: 25 examples of top-scoring boundary detection features, chosen by cross validation testing over 48 folds. Individual features are not reliable on their own – it takes at least 1,000 features to get good predictions (see text).

$$v = \sum_{i=1}^{25} e^{-\frac{\|D-D_i\|_2^2}{2\sigma^2}} v_i\,, \qquad v_i = \begin{bmatrix}\cos\theta_i \\ \sin\theta_i\end{bmatrix}$$

$$\theta = atan2(v), \qquad S = \|v\|_2$$

Where atan2 is a 4-quadrant arctangent function. σ is set to 0.25. θ, S are then used to estimate the figure/background direction at the patch in question. Estimations of all patches are added together weighted by S and smoothed by a spatial Gaussian function (positive in the figure, negative in background side). This yields a single total figure-score at each image location, where a positive score is likely to be the figure side of an object boundary.

Results Examples of boundary detection are shown in Figure 3. We used statistical testing to compare the density of boundary features in a region (10 pixels) around object boundaries compared with inside the object and on the background. Density was significantly higher around the boundaries compared with internal or external regions. In contrast, object-based segregation produced higher density in internal regions compared with boundary or external regions (1-tailed t-test, n=1200 $p < 10^{-6}$ in all comparisons).

Types And Number Of Boundary Features Psychophysical and computational studies of boundary features have suggested several types of informative boundary features, including: interposition (T-junctions), surface junctions, such as Y-junctions and arrow-junctions, and extremal edges, or folds, (Geisler *et al*. 2009, Ghose & Palmer 2010) coming from the projection of an occluding edge curving smoothly in 3D, typically creating a highlight or shadow along the curving edge.

The current study used automatically labeled object boundaries, identified by motion discontinuities.

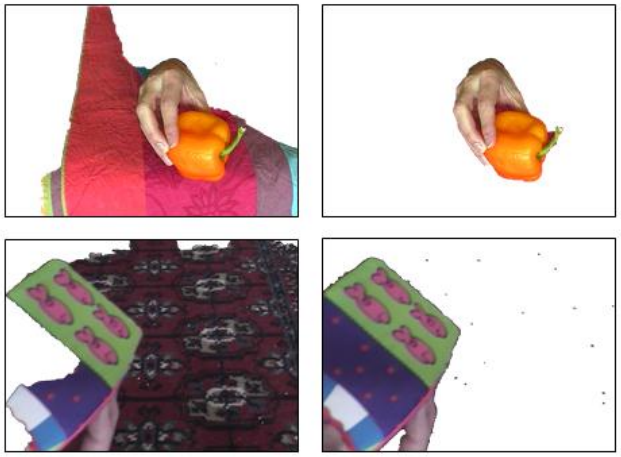

Figure 5: Object segmentation with the GrabCut algorithm. Left: Segmentation produced by the algorithm using default initialization. Right: Segmentation results with initialization by our segregation score maps.

Consequently, it became possible to extract and study a much richer set of boundary features compared with previous studies that used human annotated boundaries (Geisler *et al*. 2009, Fowlkes *et al*. 2007). The learning process produced a rich and varied set of boundary cues. Their analysis revealed the following properties. (i) Individual boundary features are probabilistic in the sense that they contribute information to the correct figure direction, but individual features are usually not definitive on their own. When training on 100,000 boundary features, the correct figure side is predicted in novel boundary features 78% of the times. (ii) Boundary features are consistent across image sets and are therefore useful for generalization to novel images. Our testing was done in 48 cross-validation folds, each time testing one movie, and excluding all movies with the same object or background from training data. (iii) There is a large set of useful boundary features, and using a restricted subset is less accurate than using the larger set. We selected the best performing features by cross-validation folds, and tested sets of different sizes, yielding 75% accuracy for 10,000 patches, 71% for 1,000, 65% for 100, 54% for 50. Nonetheless, the improvement diminishes for very large sets, suggesting that saturation may be reached at some point, and there is no need to memorize every observed feature. Exploring mechanisms of feature retention is left for future work. (iv) Among the top-scoring boundary features (examples in Fig. 4) there is a significant fraction that can be labeled 'extremal edges'. These have only recently been found to play a crucial role in human vision (Ghose & Palmer 2010), and have not been tested in infants' object segregation. Our model focuses on learning boundary features, and does not model their integration within a fully functional segregation system. To illustrate their contribution we therefore used them as input to an existing algorithm (GrabCut, Rother *et al*. 2004); results are

illustrated in Figure 5. The figure shows performance of the algorithm in its standard form (left), and the same algorithm when supplied with our object and boundary scores.

Discussion

The model demonstrates how static object segregation can be learned effectively guided by two motion based mechanisms known to be innate or early learned in infants' vision: grouping by common motion and sensitivity to motion discontinuities.

These mechanisms are used by the model for two complementary goals: common motion is used for object-based segregation, and motion discontinuities are used for learning static occlusion cues. In agreement with infants learning, the learning of object-based segregation by the model is fast, with initial sensitivity to details of the object's internal texture. It identifies well the region of the object with reduced accuracy near the boundaries. Boundary cues require more prolonged learning, but they appear to generalize broadly to novel object images. The set of useful boundary features found by the model is large and varied, including a major contribution from extremal edges, which have played a limited role in modeling so far.

The results of the study suggest a number of interesting directions for further research. In terms of infant studies, it will be of interest to test their capacity for object segregation based on extremal cues, which, to the best of our knowledge have not been tested so far. Another prediction that can be tested is whether object-based segregation by infants, which is sensitive to internal texture, will exhibit insensitivity to the object's boundary. Computationally, it will be interesting to compile a large set of useful boundary features that could be used by future segmentation algorithms. Finally, since scene segmentation in natural images is still a challenging open problem, it will be of interest to extend the current approach and examine whether following human development, by letting object segregation (including cues not considered in the current model) be guided and learned using dynamic cues, could lead to the emergence of models approaching human segregation capacities.

Acknowledgments

The work was supported by European Research Council (ERC) Advanced Grant "Digital Baby" (to S.U.).

References

Arya, S., & Mount, DN (1993). Approximate nearest neighbor queries in fixed dimensions. *ACM-SIAM*, 271–280.

Geisler, WS, Najemnik, J, Ing, AD. (2009). Optimal stimulus encoders for natural tasks. *J. Vision* 9(13), 1–16

Granrud, CE & Yonas, A. (1984) Infants' Perception of Pictorially Specified Interposition. *J. Exp. Child Psych.* 37, 500-511

Granrud CE, Yonas A, Smith IM, Arterberry ME, Glicksman ML, Sorknes AC. (1984) Infants' sensitivity to accretion and deletion of texture as information for depth at an edge. *Child Dev.* 55(4), 1630-1636.

Ghose, T, & Palmer, S. (2010) Extremal edges versus other principles of figure-ground organization. *J. Vision* 10(8)3, 1 – 17

Johnson, SP & Aslin, RN. (1996). Perception of Object Unity in Young Infants: The Roles of Motion, Depth, and Orientation. *Cog. Dev.* 11, 161-180.

Kellman, PJ, Spelke, ES, (1989). Perception of Partly Occluded Objects in Infancy. *Cog. Psych.* 15, 483-524.

Kovacs, I, Kozma, P, Feher, A, Benedek, G. (1999). Late Maturation of Visual Spatial Integration in Humans. *PNAS* 961(21), 12204-12209.

Lowe, D. (2004) Distinctive image features from scale-invariant keypoints. *Int. J. Comp. Vis.*, *60*, *2* 91-110.

McDermott, J. (2004) Psychophysics with junctions in real images. *Perception*, 33, 1101 – 1127.

Mount, D. & Arya, S. (2010) ANN: A Library for Approximate Nearest Neighbor Searching. http://www.cs.umd.edu /~mount/ANN/

Needham, A,, & Baillargeon, R. (1998). Effects of prior experience in 4.5-month-old infants' object segregation. *Inf. Behav. Dev.* 21, 1-24.

Needham, A. (2001). Object Recognition and Object Segregation in 4.5-Month-Old Infants. *J. Exp. Child Psychol*, (78), 3-24.

Needham A, Modi A., (1999). Infants' use of prior experiences with objects in object segregation: implications for object recognition in infancy. *Adv Child Dev Behav*., 27:99-133.

Riesenhuber, M. & Poggio, T. (1999), Hierarchical models of object recognition in cortex. *Nat. Neurosci*, 2(11), 1019–1025.

Ross, MG & Kaelbling, LP. (2009). Segmentation According to Natural Examples: Learning Static Segmentation from Motion Segmentation. *IEEE PAMI* 31(4), 661-676.

Rother, C., Kolmogorov, V., Blake, A. (2004) GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Transactions on Graphics*, 23 (3), 309-314.

Spelke, ES. (1990). Principles of object perception. *Cognitive Science*, 14, 29-56

Spelke, E.S, Breinlinger, K, Jacobson, K, Phillips, A. (1993) Gestalt relations and object perception: a developmental study. *Perception*, (22) 1483-1501.

Spelke, ES, von Hofsten, C, Kestenbaum, R. (1989). Object Perception in Infancy: Interaction of Spatial and Kinetic Information for Object Boundaries. *Developmental Psychology*, 25(2), 185-196.

Sun, D, Roth, S., Black, M.J. (2010). Secrets of optical flow estimation and their principles. *IEEE CVPR*, 2432-2439.

Ullman, S. Harari, D. Dorfman, N. 2012. From simple innate biases to complex visual concepts. *PNAS, 109(44)*, 18215–18220.