UNSUPERVISED VIDEO SEGMENTATION ALGORITHMS BASED ON FLEXIBLY REGULARIZED MIXTURE MODELS

Claire Launay[1]*, Jonathan Vacher*[2]*, Ruben Coen-Cagli*[1,3,4]

[1] Dept. of Systems & Comp. Biology, AECOM, Bronx, NY, USA
[2] Laboratoire des Systèmes Perceptifs, DEC, ENS, PSL University, CNRS, Paris, France
[3] Dominick P. Purpura Dept. of Neuroscience, AECOM, Bronx, NY, USA
[4] Dept. of Ophthalmology & Visual Sciences, AECOM, Bronx, NY, USA

ABSTRACT

We propose a family of probabilistic segmentation algorithms for videos that rely on a generative model capturing static and dynamic natural image statistics. Our framework adopts flexibly regularized mixture models (FlexMM) [1], an efficient method to combine mixture distributions across different data sources. FlexMMs of Student-t distributions successfully segment static natural images, through uncertainty-based information sharing between hidden layers of CNNs. We further extend this approach to videos and exploit FlexMM to propagate segment labels across space and time. We show that temporal propagation improves temporal consistency of segmentation, reproducing qualitatively a key aspect of human perceptual grouping. Besides, Student-t distributions can capture statistics of optical flows of natural movies, which represent apparent motion in the video. Integrating these motion cues in our temporal FlexMM further enhances the segmentation of each frame of natural movies. Our probabilistic dynamic segmentation algorithms thus provide a new framework to study uncertainty in human dynamic perceptual segmentation.

Index Terms— Video Segmentation, Mixture Models, Graphical Models, Optical Flows, Temporal Propagation

1. INTRODUCTION

Integrating visual features into perceptual groups and segmenting those groups from each other, is key to adaptive behavior. While there is ample literature for static image segmentation, much less is known about how groups are formed and maintained during dynamic stimulation, and Gestalt principles of perceptual organization such as the proximity rule—one of the most powerful static grouping cues—do not always generalize to temporal dynamics [2]. On the other hand, subjective experience suggests that coherent motion of objects through visual space is a powerful cue for segmentation, which has also been exploited in several algorithms [3, 4]. Therefore, it is crucial to develop theoretical frameworks that respect principles of static segmentation and extend them to dynamic inputs. Recently, some video segmentation approaches have started providing high accuracy results thanks to advances in deep learning techniques, for example by relying on a recurrent neural network to estimate spatial and temporal patterns [5] or using appearance and motion cues through a two-stream network [6] and fuse them to obtain segmentation maps. The current methods achieving state-of-the-art performances for video object segmentation [7, 8] are unsupervised deep-learning based methods using a Siamese architecture and spatial and long-term temporal context to improve time consistency. Chandra et al. [9] use spatio-temporal random fields jointly with a CNN-based per-frame segmentation to propagate temporal information. Some methods [10, 11] combine static features from adjacent frames for segmentation prediction, which can increase the computation cost [10].

Different from the approach introduced in this paper, these algorithms are not based on probabilistic models, and so cannot account for local and global uncertainty of the video segmentation. Indeed, uncertainty is a central aspect of human perception, and our long term goal is to develop an ideal observer model of human perceptual segmentation [12]. Modeling uncertainty is also important for real-life applications where e.g. crowding and occlusions can produce substantial uncertainty about segmentation. Here we propose a family of probabilistic segmentation algorithms that share grouping information across visual features in space and time, and rely on a generative model that captures static and dynamic natural image statistics. Our framework adopts flexibly regularized mixture models (FlexMM) [1], an efficient method to combine mixture distributions across different data sources. FlexMMs of Student-t distributions capture the statistics of wavelet coefficients and hidden units of deep convolutional neural networks (CNNs), and successfully segment static natural images, through uncertainty-based information sharing between hidden layers of CNNs. Here, we further extend this approach to dynamic inputs.

The contributions of this paper are the following: 1. We develop FlexMMs for temporal and spatial propagation of information, 2. We propose an efficient and flexible video segmentation algorithm, 3. We test FlexMMs on two databases to show that temporal and motion propagation throughout inference improves the segmentation of sequences of images, especially in case of large displacements.

2. STUDENT MIXTURE MODELS FOR VIDEO MODELING

Due to their ability to effectively model image statistics and pixel correlations, Gaussian mixture models (GMM) have been widely used for image segmentation [14, 15]. Yet, despite their tractability and the low number of parameters to estimate, these models are sensitive to noise and outliers. Multivariate Student-t distributions are known to be more robust to outliers than GMMs, as they are more heavily tailed. Furthermore, they capture both low-level features, such as wavelet coefficients [16], and high-level features extracted by deep CNNs [17, 18], and provide good results for static segmen-

RCC is supported by NIH (EY031166 and EY030578). JV is supported by ANR (ANR-19-NEUC-0003-01).

a. Frame 1 b. Optical flow 1→ 2 c. Frame 2 d. Optical flow 2→ 3 e. Frame 3

Fig. 1: Three frames extracted from a video in the Davis database [13] and the estimated optical flows, computed from successive frames.

tation [19, 20, 1].

Here, we also want to extract and represent dynamic features from the video to be segmented. Indeed, integrating temporal information can be crucial to maintain a consistent segmentation across the sequence. In this paper we explore two ways of integrating temporal information. First, we directly share information between the static features of successive frames, as described in the next section. Second, we use optical flows, which provide a two-dimensional representation of motion [21], its direction and speed, between successive image frames. As such, they give important cues about the position and the changes of position of objects in the scene [22], so that they may be involved in the decomposition of an image into moving objects. Figure 1 illustrates OFs obtained from adjacent frames of a sequence, where the direction and speed of motion are represented by hue and saturation, respectively. Statistics of OFs in natural movies have not been studied extensively, but it is known that they display sharp edges and follow a heavy-tail distribution, two properties similar to static features of natural images. Roth [23] also showed that OFs histograms are well captured by Student-t distributions. As one can observe in Figure 1, OFs also convey precise boundary information for moving objects and occlusions that can be essential cues in scenes with low-visibility conditions. Thus, OFs can be represented by a generative model using mixtures of multivariate Student-t distribution and estimating the mixing parameters of this model, giving a motion segmentation map, and then enhance video segmentation.

3. MODEL FOR DYNAMIC SEGMENTATION

The model presented here builds on FlexMM [1], and extends it to capture sequences of images and spatially segment them. Consider an image domain of size N and a number of frames T. We assume that each pixel n in frame t is characterized by a feature vector $x_{n,t}$ associated with a random vector $X_{n,t}$. Given a number K of segments in the sequence, our goal is to link each pixel to a class, modeled by the random variable $C_{n,t}$, so as to obtain probabilistic segmentation maps $(\mathbf{p}_{n,t,k} = \mathbb{P}(C_{n,t} = k))_{n,k}$ for every frame t of the sequence. The original formulation of these FlexMMs takes advantage of the descriptive strength and tractability of finite mixture models. FlexMMs extend finite mixtures to impose a spatial and multi-source transfer of information by over-parametrizing mixing probabilities $\mathbf{p}$ and adding a specific prior on them: $\mathbb{P}_{X_n,\mathbf{P}_n|\mathbf{A}}(x, \mathbf{p}_n|\mathbf{a}) = \mathbb{P}_{\mathbf{P}_n}(\mathbf{p}_n) \sum_{k=1}^{K} p_{n,k} \mathbb{P}_{X^{(k)}}(x; a_k)$, where for all $n \in \{1, \dots, N\}$, $k \in \{1, \dots, K\}$, $0 \leq p_{n,k} \leq 1$, $\sum_{k=1}^{K} p_{n,k} = 1$ and $\mathbb{P}_{X^{(k)}}$ represents the distribution of the feature random vector given its class is k. We choose Student-t distributions with parameters a_k to model these feature vectors. We also assume that the mixing probabilities follow a Dirichlet distribution $\mathbf{P}_n \sim \mathcal{D}(\mathbf{B}_n)$, where parameter $\mathbf{B}_n$ is a random vector whose distribution is linearly determined by the classes of the other pixels and the other frames of the sequence $\forall k \in \{1, \dots, K\}$,

$$B_{n,t,k} = \sum_{n',t'} \lambda_{n,t,k}(n', t') \mathbb{1}_k(C_{n',t'}) - \mathbb{1}_k(C_{n,t}) + 1, \qquad (2)$$

with $(\lambda_{n,t,k}(n', t'))_{n' \in \{1,\dots,N\}, t' \in \{1,\dots,T\}}$ a set of $N \times T$ weights that are specifically determined in function of the current pixel, frame and segment. Under these assumptions, the mixing probabilities, meaning the probabilistic segmentation maps, and the parameters of the Student-t distribution are inferred using an expectation-maximization (EM) algorithm. At each iteration i, the mixing probabilities are updated as a linear combination of the current posterior probabilities $\forall k \in \{1, \dots, K\}, \forall n \in \{1, \dots, N\}, \forall t \in \{1, \dots, T\}$,

$$p_{n,t,k}^{(i+1)} = \sum_{n',t'} \omega_{n,t,k}(n', t') \tau_{n',t',k},$$

with $\tau_{n,t,k}^{(i)} = \mathbb{P}\left(C_{n,t} = k \,|\, X_{n,t} = x_{n,t}, \boldsymbol{p}^{(i)}, \boldsymbol{a}^{(i)}\right)$ the estimated posterior probabilities at iteration i and $\boldsymbol{\omega}_{n,t,k}$ weights given by $\lambda_{n,t,k}$ and a normalization constant. These weights can be chosen freely and adapted to the sequence of images, the current frame t or current pixel n. This model allows for spatial and temporal smoothing, i.e. the propagation of information between pixels and also between frames. Figure 3 presents a simplified graph representation of our generative model for a sequence of three frames made of two pixels. The first strength of this model is its tractability: at each iteration of the EM algorithm, the update rule of the mixing probabilities $p_{n,t,k}$ is linear. A second advantage of this model is its flexibility, because it is possible to entirely define how information is combined and spread by choosing specific weights $\lambda_{n,t,k}$. In what follows, we compare several options for these weights and the way the posterior probabilities are combined during inference. Our first model, called *Temp-prop* in the rest of the paper, combines spatial and temporal local class assignments, as shown in Figure 3. We define the weights $\lambda_{n,t,k}$ so that the updated mixing probability is a linear combination using three spatial Gaussian kernel, each one averaging locally the posterior maps of the previous, current and next frames. The update rule is given in Equation (1), where G is a 2D Gaussian kernel, and $m_{n,t,k}^{(i)} = G * \tau_{.,t,k}^{(i)}$, and $s_{n,t}^{(i)\,2} = \frac{\sum_{k=1}^{K} G*{\tau_{.,t,k}^{(i)}}^2(n) - {m_{n,t,k}^{(i)}}^2}{K(1 - G*G(0))}$ can be seen as the spatial mean and the variance of the posterior map around pixel n, at frame t, for segment k. Therefore, for a given frame, posterior maps that locally have a high variance (high uncertainty) will influence relatively less the mixing probabilities of the neighboring frames. Our second model, called *Temp+OF prop* similarly combines the posterior maps from the current frame and from optical flows from the previous and the next frames, incorporating strong motion clues in the segmentation process.

At the end of the EM algorithm, we obtain probabilistic segmentation maps $(p_{n,t,k})_n$ for each frame $t \in \{1, \dots, T\}$ and each segment $k \in \{1, \dots, K\}$ (Figure 3). At each position, the most-probable segment is selected to obtain a segmentation map as shown in Figure 3 (bottom row), each color corresponding to one segment.

$$p_{n,t,k}^{(i+1)} = \frac{{s_{n,t}^{(i)}}^2 {s_{n,t+1}^{(i)}}^2 m_{n,t-1,k}^{(i)} + {s_{n,t-1}^{(i)}}^2 {s_{n,t+1}^{(i)}}^2 m_{n,t,k}^{(i)} + {s_{n,t-1}^{(i)}}^2 {s_{n,t}^{(i)}}^2 m_{n,t+1,k}^{(i)}}{{s_{n,t}^{(i)}}^2 {s_{n,t+1}^{(i)}}^2 + {s_{n,t-1}^{(i)}}^2 {s_{n,t+1}^{(i)}}^2 + {s_{n,t-1}^{(i)}}^2 {s_{n,t}^{(i)}}^2} \quad (1)$$

Fig. 2: Update rule for the mixing probabilities $\mathbf{p}$ during iteration i of the EM algorithm. See end of Section 3 for explanations.

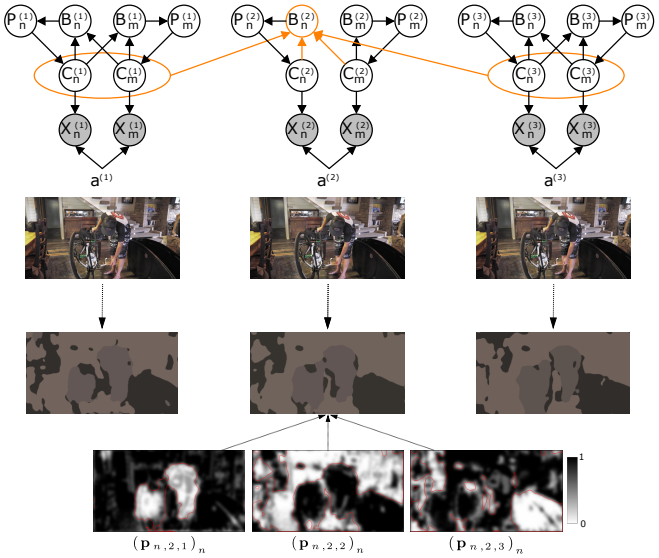


Fig. 3: Graphical model for our *Temp-prop* model: the prior of the mixing probabilities depends on the class labels of the neighboring pixels and of the previous and next frames. Each frame is associated with one mixture model. All segmentation maps are estimated simultaneously by the EM algorithm. The last row displays the probabilistic segmentation maps associated with the second frame of the video segmented into three segments.

4. RESULTS

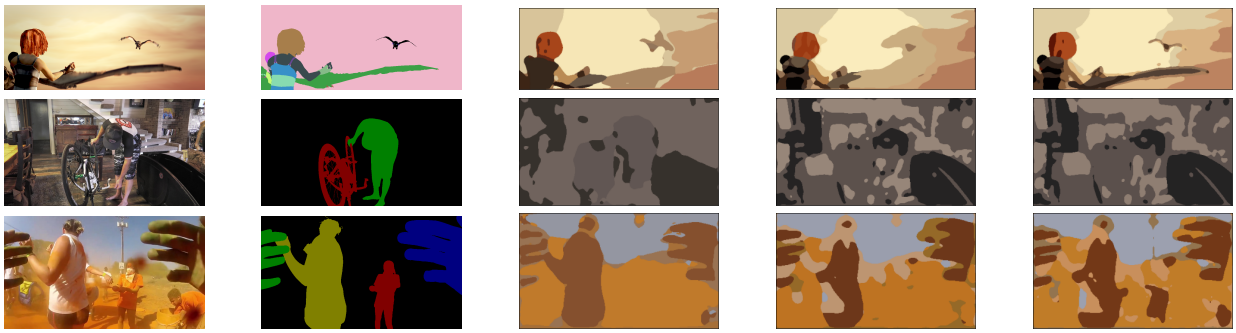
Implementation details. In our model, the observations $(x_{n,t})_{n,t}$ are image features extracted by the pre-trained deep network VGG19 [24]. In the two implementations presented here (*Temp prop* and *Temp + OF prop*), we limit ourselves to the features of the first layer of the network. In the second model, we also use features of optical flows estimated for each pair of frames to influence the segmentation of the sequence. To estimate each optical flow from a pair of frames, we used a method called PWC-net [25], an algorithm based on CNNs that offers a good compromise between quality and efficiency. We compare our two model implementations with a model that uses the same static features as *Temp prop*, ie. layer 1 of VGG19, but ignores temporal propagation. We compare with this model to evaluate the effects of time propagation on image segmentation and on temporal consistency. In addition, we also compare to the static model of [1], handling each frame independently and propagating class information between hidden units of deep CNNs. This hierarchical model uses features from the 16 first layers of VGG19 and slightly outperforms comparable unsupervised algorithms [26, 27, 28], though it does not match state-of-the-art and deep learning based algorithms. Our main algorithms TempProp and Temp+OF run in 201.8s and 395.7s for a 40-frames sequence, 240x416 pixels on a single Intel Xeon core (E5-2680), 7Gb RAM. The hierarchical models are an order of magnitude slower than those using only one layer of features, indicating that the number of features is the key determinant of computational cost.

Qualitative results. We compared these algorithms using two databases: the DAVIS database [13], a widely used database for natural video segmentation, and the MPI-Sintel database [29], gathering synthetic videos with complex motion patterns used to evaluate OF estimation and video segmentation algorithms. Figure 4 presents the segmentation maps for three sequences, produced by *Temp prop*, *Temp + OF prop* and the hierarchical model of [1]. First, temporal propagation across the iterations of the EM algorithm, used in our two models, seems to increase spatial smoothing. In the three sequences, both models tend to create larger and more connected segments, compared to the hierarchical model which segments each frame independently. This effect is even more pronounced for the model *Temp + OF prop* that propagates motion estimation. These examples show how this model retrieves some object-level information thanks to their motion, while the model based only on temporal propagation and the hierarchical model seem more sensitive to illumination changes and shadows (first and third row). Similarly, in a complex scene such as the second row, OFs extract meaningful regions of the scene that could hardly be retrieved by the other two models. However, the model *Temp + OF prop* tends to gather moving parts that don't belong to the same objects (second row) and to omit static objects in the scene (the second person, third row). This highlights the issue of dealing with inconsistencies between motion estimation and static features and how to prioritize between these conflicting cues [2]. Note also that better spatial smoothing and objects integrity come at the price of lower resolution in the resulting segmentation maps. Additional videos comparing these models are available online[1].

Quantitative results. To quantify the performance of these methods, we used two common evaluation metrics. First, as our main focus is capturing temporal consistency of segmentation maps across the sequence, we applied a temporal stability metric that uses OFs to compute the overlap between the warped segmentation map of the previous frame and current frame. This metric does not use the ground truth segmentation, but only assesses temporal consistency of the model segmentation. In addition, we also tested if temporal integration helps static segmentation as suggested by Fig. 4, using the Jaccard index that measures the overlap between regions in the estimated and the true segmentation maps for every frame of the sequence independently. This is a measure of how well the static segmentation of individual frames matches the ground truth. The scores for both databases [29, 13] are reported in Figure 5.

Temporal stability across frames is particularly enhanced for both temporal models compared to the static segmentation algorithm (model *Indep L1* using the first layer of VGG19). We found that the initialization has also a strong impact on temporal consistency, as we set the first frame initialization to influence the subsequent ones. When the initialization is shared across frames but the estimation process is still performed independently (*Indep SI L1*), temporal stability is closer to the temporal models, but still lower indicating that temporal propagation of class information during inference is essential to capture temporal stability. These two new variations perform slightly poorer on the MPI-Sintel database, where large displace-

[1] https://claunay.github.io/videoFlexMM.html



a. Original frame b. Ground truth c. Temp + OF prop d. Temp prop e. Hierar. model

Fig. 4: Segmentation of natural and synthetic videos using both models presented in Section 3 and the hierarchical model of [1] (combining image features extracted from 16 layers of a VGG19 network).

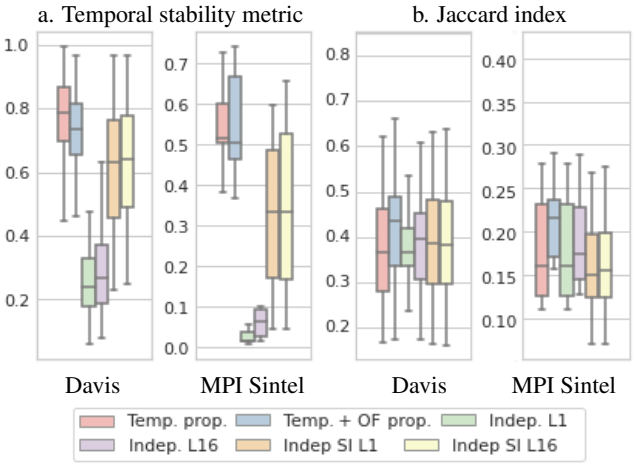


Fig. 5: Performances of the two models presented in this paper, compared with variants of the static model for image segmentation of [1], using one (Indep. L1) or 16 layers (Indep. L16) of VGG19. The temporal stability metric charts also compares these models with variations of the hierarchical model. The horizontal bar represents the median, shaded boxes represent the quartiles of the dataset while the errors bars show the rest of the distribution, except for outliers.

ments occur. The Jaccard index computed for both databases shows that the model using OFs seems to slightly outperform the model using only temporal propagation (*Temp prop*) and their equivalent static algorithm (*Indep L1*). For a stronger test, we also compared our models with the hierarchical model using 16 layers of VGG19. Notice that this model takes advantage of the image features from the 16 layers of a deep CNN, while the two models presented here only use features from the first layer of this network. We retrieve the same conclusions as with the model *Indep L1*, that is the two models presented in this paper provide a significant improvement of temporal stability when the initialization is independent and slightly increase the Jaccard index.

5. DISCUSSION

We have introduced a new method for unsupervised video segmentation based on probabilistic mixture models. Modeling videos using a mixture model and estimating the parameters of the model are complex tasks, due to the high dimensionality of the feature space. To offer an alternative to a spatio-temporal mixture model, which would be too heavy to infer, we extended FlexMMs [1] to share information between mixture models associated with each frame of the sequence. In the resulting family of models, segmentation information is spread across space and time in an adjustable fashion. We have shown that the propagation of information between layers during the estimation process improves temporal consistency. In addition, the approach further enhances spatial smoothing and object retrieval compared to the initial static algorithms, even when deeper layers of CNN are used and especially when using optical flows.

Note that in some cases integrating information across frames is not the best strategy, e.g. if there is a large change such as a scene cut or a large eye movement. Our framework has the ability to adapt to these changes by defining the appropriate weights λ (Eq (2)) so that we can insert a change-detection mechanism to turn off temporal integration when necessary. As observed in Figure 4 (third row), integrating motion and temporal cues can also be at the expense of the detection of static objects. This observation shows the necessity to properly fuse motion and appearance cues in video segmentation algorithms and also to further explore how human perception combines dynamic and static features, in particular when those cues are in conflict [2, 30]. The flexibility of this new family of models enables to adapt the segmentation to this future analysis. A key property of the model is that it produces probabilistic segmentation maps per frame. Computationally, we take advantage of this measure of uncertainty (or precision) during inference, by averaging the probabilistic segmentation maps of each frame weighted by their local precision. In future work, this knowledge about the uncertainty in the resulting segmentation can be compared to human perceptual segmentation data, to test the hypothesis that human variability in video segmentation may reflect perceptual uncertainty [12].

Our framework constitutes a step forward in developing ideal observer models for the segmentation of dynamic natural inputs. To enhance the biological foundation of our model, in future work we will further examine different temporal features, for instance by using a model of neuronal responses in visual-cortical area MT [31] or biologically inspired algorithms to compute OFs [22].

6. REFERENCES

[1] Jonathan Vacher, Claire Launay, and Ruben Coen-Cagli, "Flexibly regularized mixture models and application to image segmentation," *Neural Networks*, 2022.

[2] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger von der

Heydt, "A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization.," *Psychological bulletin*, vol. 138, no. 6, pp. 1172, 2012.

[3] Daniel Cremers and Stefano Soatto, "Motion competition: A variational approach to piecewise parametric motion segmentation," *International Journal of Computer Vision*, vol. 62, no. 3, pp. 249–265, 2005.

[4] P. Ochs, J. Malik, and T. Brox, "Segmentation of moving objects by long term video analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 6, pp. 1187 – 1200, Jun 2014, Preprint.

[5] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao, "Zero-shot video object segmentation via attentive graph neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9236–9245.

[6] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang, "Segflow: Joint learning for video object segmentation and optical flow," 09 2017.

[7] Xiankai Lu, Wenguan Wang, Chao Ma, Jianbing Shen, Ling Shao, and Fatih Porikli, "See more, know more: Unsupervised video object segmentation with co-attention siamese networks," 06 2019.

[8] Sucheng Ren, Wenxi Liu, Yongtuo Liu, Haoxin Chen, Guoqiang Han, and Shengfeng He, "Reciprocal transformations for unsupervised video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15455–15464.

[9] Siddhartha Chandra, Camille Couprie, and Iasonas Kokkinos, "Deep spatio-temporal random fields for efficient video segmentation," 12 2018.

[10] Raghudeep Gadde, Varun Jampani, and Peter V Gehler, "Semantic video cnns through representation warping," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4453–4462.

[11] Evan Shelhamer, Kate Rakelly, Judy Hoffman, and Trevor Darrell, "Clockwork convnets for video semantic segmentation," 08 2016.

[12] Jonathan Vacher, Pascal Mamassian, and Ruben Coen-Cagli, "An ideal observer model to probe human visual segmentation of natural images," *arXiv preprint arXiv:1806.00111*, 2018.

[13] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Computer Vision and Pattern Recognition*, 2016.

[14] Haim Permuter, Joseph Francos, and Ian Jermyn, "A study of gaussian mixture models of color and texture features for image classification and segmentation," *Pattern Recognition*, vol. 39, no. 4, pp. 695–706, 2006, Graph-based Representations.

[15] Thanh Minh Nguyen and Q. M. Jonathan Wu, "Robust student's-t mixture model with spatial constraints and its application in medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 31, no. 1, pp. 103–116, 2012.

[16] Martin J Wainwright and Eero P Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Advances in neural information processing systems*, 2000, pp. 855–861.

[17] Luis G Sanchez-Giraldo, Md Nasir Uddin Laskar, and Odelia Schwartz, "Normalization and pooling in hierarchical models of natural images," *Current opinion in neurobiology*, vol. 55, pp. 65–72, 2019.

[18] Jonathan Vacher, Aida Davila, Adam Kohn, and Ruben Coen-Cagli, "Texture interpolation for probing visual perception," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[19] Giorgos Sfikas, Christophoros Nikou, and Nikolaos Galatsanos, "Robust image segmentation with mixtures of student's t-distributions," in *2007 IEEE International Conference on Image Processing*. IEEE, 2007, vol. 1, pp. I–273.

[20] Dibyendu Mukherjee and Q.M. JonathanWu, "Real-time video segmentation using student's t mixture model," *Procedia Computer Science*, vol. 10, pp. 153–160, 2012, ANT 2012 and MobiWIS 2012.

[21] Berthold K.P. Horn and Brian G. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, no. 1, pp. 185–203, 1981.

[22] Cornelia Beck, Thilo Ognibeni, and Heiko Neumann, "Object segmentation from motion discontinuities and temporal occlusions–a biologically inspired model," *PLOS ONE*, vol. 3, no. 11, pp. 1–14, 11 2008.

[23] Stefan Roth and Michael J Black, "On the spatial statistics of optical flow," *International Journal of Computer Vision*, vol. 74, no. 1, pp. 33–50, 2007.

[24] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8934–8943.

[26] H. Steinhaus, "Sur la division des corps matériels en parties," *Bulletin de l'Académie Polonaise des Sciences*, vol. Cl. III — Vol. IV, no. 12, pp. 801–804, 1956.

[27] Tian Zhang, Raghu Ramakrishnan, and Miron Livny, "Birch: an efficient data clustering method for very large databases," *ACM sigmod record*, vol. 25, no. 2, pp. 103–114, 1996.

[28] Dorin Comaniciu and Peter Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[29] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *European Conf. on Computer Vision (ECCV)*, A. Fitzgibbon et al. (Eds.), Ed. Oct. 2012, Part IV, LNCS 7577, pp. 611–625, Springer-Verlag.

[30] Maggie Shiffrar, Xiaojun Li, and Jean Lorenceau, "Motion integration across differing image features," *Vision research*, vol. 35, no. 15, pp. 2137–2146, 1995.

[31] Eero P. Simoncelli and David J. Heeger, "A model of neuronal responses in visual area mt," *Vision Research*, vol. 38, no. 5, pp. 743–761, 1998.