On Scene Segmentation and Histograms-Based Curve Evolution

Amit Adam, Ron Kimmel, *Fellow*, *IEEE*, and Ehud Rivlin

Abstract—We consider curve evolution based on comparing distributions of features, and its applications for scene segmentation. In the first part, we promote using cross-bin metrics such as the Earth Mover's Distance (EMD), instead of standard bin-wise metrics as the Bhattacharyya or Kullback-Leibler metrics. To derive flow equations for minimizing functionals involving the EMD, we employ a tractable expression for calculating EMD between one-dimensional distributions. We then apply the derived flows to various examples of single image segmentation, and to scene analysis using video data. In the latter, we consider the problem of segmenting a scene to spatial regions in which different activities occur. We use a nonparametric local representation of the regions by considering multiple one-dimensional histograms of normalized spatiotemporal derivatives. We then obtain semisupervised segmentation of regions using the flows derived in the first part of the paper. Our results are demonstrated on challenging surveillance scenes, and compare favorably with state-of-the-art results using parametric representations by dynamic systems or mixtures of them.

Index Terms—Segmentation, Earth Mover's Distance, curve evolution, scene analysis.

1 Introduction

1.1 Motivation and Overview

Motivated by surveillance applications, our goal is to identify, in a given scene, different regions where different types of activities take place. Rather than dealing with image segmentation, we use dynamics-based observations, extracted from a video stream from the scene, in order to segment the scene into various spatial regions.

In this paper, we use a curve evolution variational framework for segmentation. The flow fields driving the curves are based on the distributions of features in the inner and outer regions bounded by the curves. Therefore, in the first part of this paper, we consider the general problem of histogram-based curve evolution. We derive novel flow fields to guide the evolution process, based on using the Earth Mover's Distance (EMD) [29] for measuring the dissimilarity between two histograms. We apply these flows to various examples and discuss their limitations.

In the second part of this paper, we return to the problem of activity-based scene segmentation. We model regions by histograms of features and apply the relevant flow fields for obtaining semisupervised region segmentation.

1.2 Related Work

1.2.1 Curve Evolution Using Region Statistics

Active contours techniques were originally based on flow fields derived by integrating region boundary image data—such as edge strength—with a contour regularization term [4]. Early papers using region data to drive the curve are, for example, the papers by Ronfard [27], Zhu and Yuille [33], and Chan and Vese [7]. A review of level-set methods based on region statistics was recently presented by Cremers et al. [10]. Jehan-Besson et al. [19] present a general scheme for deriving flow fields that minimize a functional constituting of a regional integral of certain descriptors which themselves depend on the region. The descriptors integrated within a region may also depend on the region—for example, a probability density on the region. We adopt this framework in order to derive the flow fields presented in this paper.

More specifically, histograms were used as the regional descriptors in various papers. In Freedman and Zhang's work [16], the Kullback-Leibler and Bhattacharyya distances between the sample density inside the region and a template density are used as the minimization objective. Michaelovich et al. [24] maximize the Bhattacharyya distance between the density inside the region and the density outside of the bounded region. Aubert et al. [1] compare histograms again by using the Kullback-Leibler distance and the Hellinger distance which is equivalent to the Bhattacharyya metric. A disadvantage these measures have is their sensitivity to quantization: Some bin differences may be due to the quantization of measurements into bins and not to differences in the actual distributions.

The first contribution of this paper is the derivation of gradient flow fields with respect to the Earth Mover's Distance metric. We first derive a simple and efficient expression for the EMD on one-dimensional histograms. Then, we use this tractable expression to derive the corresponding gradient flow fields. A related derivation has also been presented by Chan et al. [6]. We present an intuitive proof of the tractable EMD expression, and use slightly different functionals for which we derive the corresponding flows.

Since the flow fields that drive the curves implement gradient descent for minimizing a functional, the success of the method relies on the given initial contour from which we start the process. A second contribution is a method for obtaining a rough initial segmentation, using again histogram descriptors. We limit our use of histograms to rectangular regions, which allows us to exploit the efficiency of the integral histogram data structure [26] in this initialization stage.

1.2.2 Segmentation of Dynamic Scenes

There has been a lot of research on various segmentation tasks using spatiotemporal data. A lot of research considered segmentation of objects using their motion, e.g., [11], [18], [20], [23]. Another issue which received a lot of attention is temporal segmentation of the video, e.g., [32]. Examples of techniques employed on spatiotemporal data are Gaussian mixture models [17], mean-shift [12], and spectral clustering [15]. However, relatively few papers addressed the problem we are considering here—namely *scene segmentation using spatiotemporal data*.

The works most directly related to our problem are those inspired by the dynamic texture models introduced by Doretto et al. [13]. There, linear dynamic systems were shown to model certain types of videos of dynamic phenomena such as fire, flowing water, smoke, etc. Following that framework, the works in [14], [9] have considered the problem of spatial segmentation of a scene using a video stream of the scene. More recently, Chan and Vasconcelos [5] considered mixtures of dynamic textures and have shown their applicability to a wider range of videos including traffic and pedestrian videos. Our framework is different from these previous efforts in that our approach is purely nonparametric. In addition, we go beyond the smoke/fire/water type of videos and demonstrate the applicability of our approach to surveillance-type videos of various scenes.

1.3 Summary of Contributions

We summarize the main contributions of this paper:

1. Derivation of gradient flows for minimizing functionals that use the EMD. A closely related result has appeared (independently) in [6]. We provide an intuitive proof for the tractable EMD expression which allows the flows to be derived.

• *The authors are with the Department of Computer Science, Technion—Israel Institute of Technology, Haifa 32000, Israel. E-mail: {amita, ron, ehudr}@cs.technion.ac.il.*

Manuscript received 24 May 2008; revised 10 Nov. 2008; accepted 6 Jan. 2009; published online 13 Jan. 2009.
Recommended for acceptance by S.-C. Zhu.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-05-0304.
Digital Object Identifier no. 10.1109/TPAMI.2009.21.

2. Consideration of the question of initializing the curve evolution process. We exploit the integral histogram data structure and use standard clustering techniques on a set of local histograms to automatically obtain a meaningful initial contour.
3. The use of nonparametric models of different regions to represent dynamic scenes, in contrast to modeling by linear dynamic systems [14], [9], [5]. Our nonparametric model does not involve the calculation of optical flow in contrast with [21].
4. Demonstration of our approach on cluttered real-life scenes with lots of human activity, unlike the scenarios analyzed in [14], [9], [21].

2 The EMD and a Tractable Expression for One-Dimensional Densities

Let $\{p_1,\ldots,p_N\}$ and $\{q_1,\ldots,q_N\}$ be the two histograms, where we name the bins as $1,2,\ldots,N$. In this paper, all histograms are normalized (sum up to unity). Assume that the cost of moving a unit of probability mass from bin i to bin j is c_{ij} ($c_{ii}=0$). The idea behind the EMD [29] is to find the cheapest way to transfer the probability mass from $\{p_1,\ldots,p_N\}$ to $\{q_1,\ldots,q_N\}$. The cost of this cheapest possible transfer measures the distance between the two histograms.

Let us now limit ourselves to cases in which the histograms are one-dimensional. In this case, there is a unique "path" from every bin i to every bin j and it necessarily passes through other bins in between. The unique path between every two bins significantly constrains the possible flows and leads to the following:

Claim. *Let $P=\{p_1,\ldots,p_N\}$ and $Q=\{q_1,\ldots,q_N\}$ be discrete probability distributions on N bins. Let $\{P_1,\ldots,P_N=1\}$ and $\{Q_1,\ldots,Q_N=1\}$ be the corresponding cumulative distribution functions (CDFs): $P_i=p_1+\cdots+p_i, Q_i=q_1+\cdots+q_i$.*

Assume that the cost structure is $c_{ij}=|i-j|$. Then,

$$\mathbf{EMD}(P,Q)=\sum_{i=1}^{N}|P_i-Q_i|. \tag{1}$$

Proof. First suppose that $P_1\geq Q_1, P_2\geq Q_2,\ldots,P_N=Q_N=1$. The first bin currently contains p_1 mass, and needs to contain $q_1\leq p_1$ mass. Therefore, we have to move p_1-q_1 mass out, necessarily to bin 2. Now, we have $p_2+p_1-q_1$ mass in bin 2. We need leave there only q_2 mass. By the assumption $P_2\geq Q_2$, we have $p_2+p_1-q_1\geq q_2$ so we have to move $p_2+p_1-q_1-q_2=P_2-Q_2$ mass out of this bin, necessarily to bin 3.

Note that the cost for the mass moving we have done so far is $|P_1-Q_1|+|P_2-Q_2|$. We continue this way and obtain, in this specific case, that the only feasible flow that transforms $\{p_1,\ldots,p_N\}$ to $\{q_1,\ldots,q_N\}$ indeed costs $\sum_i|P_i-Q_i|$.

Thus, we have proven the claim for the case where the two CDFs do not intersect—i.e., one CDF is always greater or equal to the other CDF. In the general case where the CDFs intersect, we apply the same argument over each of the segments where one CDF is larger than the other. □

Note that our proof relies on the specific cost structure that we chose. Also note that the standard Kolmogorov-Smirnov statistic is defined by $KS(P,Q)=\max_i|P_i-Q_i|$. The claim we proved shows that the EMD distance replaces the L_∞ norm between the CDFs by the L_1 norm.

3 Curve Evolution Based on EMD

We now use the simplified form of the EMD between two distributions derived in (1) to obtain gradient flows for functionals involving the EMD.

3.1 Maximal-Discrepancy Functional

We assume that our features are one-dimensional (e.g., intensities, single-variable derivatives, edge orientations, etc.) and that we characterize a region by the distribution of features in the region. Let us represent a distribution in a nonparametric form by using a histogram with $z_1<z_2<\cdots<z_N$ being the maximal values in each bin.

Let Ω be a closed curve in the image plane. Let

$$\mathbf{CDF}_R(z_i)=\text{Prob}(\text{feature value in region } R\leq z_i), \tag{2}$$

where $R=\{\text{in},\text{out}\}$ is one of the two regions separated by Ω. We would like to find the curve Ω and its associated inside and outside regions, such that the discrepancy between the features' distributions inside and outside is maximal. We will measure the discrepancy by the EMD

$$\mathbf{EMD}(\Omega)=\mathbf{EMD}(P_{in},P_{out})=\sum_{i=1}^{N}|\mathbf{CDF}_{in}(z_i)-\mathbf{CDF}_{out}(z_i)|. \tag{3}$$

Following [19], we represent our functional (3) as a sum of (absolute value of) expressions of the form

$$\int\int_{\Omega_{out}}k^{out}_{z_i}(x,y,\Omega_{out})dxdy+\int\int_{\Omega_{in}}k^{in}_{z_i}(x,y,\Omega_{in})dxdy. \tag{4}$$

Let us define $V(z_i,\Omega)=\text{CDF}_{in}(z_i)-\text{CDF}_{out}(z_i)$ and

$$T_z(x,y)=\begin{cases}1, & I(x,y)\leq z\\ 0, & I(x,y)>z,\end{cases} \tag{5}$$

where $I(x,y)$ is the feature value at (x,y). Then,

$$\mathbf{CDF}_R(z_i)=\mathbf{Prob}(I(x,y)\leq z_i|(x,y)\epsilon R)=\frac{\int\int_R T_{z_i}(x,y)dxdy}{\int\int_R 1dxdy}. \tag{6}$$

Denote the denominator (using the same notation as in [19]) by $A_R=G_1^R=\int\int_R 1dxdy$ and define

$$k^{in}_{z_i}(x,y,\Omega)=g^{in}(x,y,G_1^{in})=\frac{T_{z_i}(x,y)}{G_1^{in}}.$$

Then, we have

$$\text{CDF}_{in}(z_i)=\int\int_{\Omega_{in}}k^{in}_{z_i}(x,y,\Omega)dxdy=\int\int_{\Omega_{in}}\frac{T_{z_i}(x,y)}{G_1^{in}}dxdy,$$

and in a similar fashion,

$$-\text{CDF}_{out}(z_i)=\int\int_{\Omega_{out}}k^{out}_{z_i}(x,y,\Omega)dxdy=\int\int_{\Omega_{out}}-\frac{T_{z_i}(x,y)}{G_1^{out}}dxdy.$$

Following [19], the flow that minimizes

$$V(z_i,\Omega)=\int\int_{\Omega_{out}}k^{out}_{z_i}(x,y,\Omega_{out})dxdy+\int\int_{\Omega_{in}}k^{in}_{z_i}(x,y,\Omega_{in})dxdy$$

is given by

$$\vec{F}=\left[k^{in}_{z_i}-k^{out}_{z_i}+A_1^{in}H_1^{in}-A_1^{out}H_1^{out}\right]\vec{N}, \tag{7}$$

where $\vec{N}$ is the inward normal to the curve, and A_1^{in},H_1^{in} are as follows:

$$\begin{aligned}A_1^{in}&=\int\int_{\Omega_{in}}\frac{\partial g^{in}}{\partial G_1^{in}}(x,y,G_1^{in})dxdy=\int\int_{\Omega_{in}}\frac{-T_{z_i}(x,y)}{{G_1^{in}}^2}dxdy\\&=-\frac{1}{{G_1^{in}}^2}\int\int_{\Omega_{in}}T_{z_i}(x,y)dxdxy\end{aligned} \tag{8}$$

and $H_1^{in}\equiv 1$ (G_1^{in} integrates 1 over Ω_{in}). Similarly, $A_1^{out}=\frac{1}{{G_1^{out}}^2}\int\int_{\Omega_{out}}T_{z_i}(x,y)dxdxy$.

Plugging everything in (7), and writing A_{in}, A_{out} instead of G_1^{in}, G_1^{out}, we obtain

$$\vec{F}_{z_i}(x,y) = \left[\frac{T_{z_i}(x,y)}{A_{in}} + \frac{T_{z_i}(x,y)}{A_{out}} - \frac{C_{in}}{{A_{in}}^2} - \frac{C_{out}}{{A_{out}}^2}\right]\vec{N}, \tag{9}$$

$$\begin{aligned} C_{in} &= \int\int_{\Omega_{in}} T_{z_i}(x,y)dxdxy \\ &= \text{number of pixels inside, where } I(x,y) \le z_i, \end{aligned} \tag{10}$$

$$\begin{aligned} C_{out} &= \int\int_{\Omega_{out}} T_{z_i}(x,y)dxdxy \\ &= \text{number of pixels outside, where } I(x,y) \le z_i. \end{aligned} \tag{11}$$

This flow minimizes $V(z_i,\Omega) = \mathrm{CDF}_{in}(z_i) - \mathrm{CDF}_{out}(z_i)$. Since we want to maximize $\mathrm{EMD}(\Omega) = \sum_{i=1}^{N} |\mathrm{CDF}_{in}(z_i) - \mathrm{CDF}_{out}(z_i)|$, the resulting flow is computed as follows:

Algorithm 1.

1. For every threshold z_i, compute

$$s_i = \mathtt{sign}[\mathrm{CDF}_{in}(z_i) - \mathrm{CDF}_{out}(z_i)] = \begin{cases} 1, & \mathrm{CDF}_{in}(z_i) \ge \mathrm{CDF}_{out}(z_i) \\ -1, & \mathrm{CDF}_{in}(z_i) < \mathrm{CDF}_{out}(z_i). \end{cases} \tag{12}$$

2. Compute

$$F_{z_i}(x,y) = \left[T_{z_i}(x,y)\left(\frac{1}{A_{in}} + \frac{1}{A_{out}}\right) - \frac{C_{in}}{{A_{in}}^2} - \frac{C_{out}}{{A_{out}}^2}\right], \tag{13}$$

where C_{in}, C_{out} are as defined in (10), (11) and A_{in}, A_{out} are the areas of the inside and outside regions determined by Ω, respectively.

3. The overall flow field is then given by

$$T(x,y) = \sum_{z_i}(-s_i F_{z_i}(x,y))\vec{N}, \tag{14}$$

where $\vec{N}$ is the inward normal.

3.2 Match-to-Template Functional

Another useful criterion is to measure how close the distribution inside the curve is to a given template distribution. Let H be the template histogram which is fixed and independent of the curve Ω. Let $J(\Omega) = \mathrm{EMD}(\text{distribution inside } \Omega, H)$ be the functional. Then, by the claim in Section 2, we have

$$J(\Omega) = \sum_{i=1}^{N} |\mathrm{CDF}_{in}(z_i) - \mathrm{CDF}_{H}(z_i)|. \tag{15}$$

In a similar fashion to Algorithm 1, the overall flow for minimizing the discrepancy with respect to a given template goes as follows:

Algorithm 2.

1. For every threshold z_i, compute $\alpha_i = \mathrm{CDF}_H(z_i)$ and $s_i = \mathtt{sign}[\mathrm{CDF}_{in}(z_i) - \alpha_i]$.
2. Compute

$$F_{z_i}(x,y) = \left[T_{z_i}(x,y)\frac{1}{A_{in}} - \frac{C_{in}}{{A_{in}}^2}\right], \tag{16}$$

where C_{in} is as defined in (10) and A_{in} is the area of the inside region determined by Ω.

3. The overall flow field is

$$T(x,y) = \sum_{z_i}(s_i F_{z_i}(x,y))\vec{N}, \tag{17}$$

where $\vec{N}$ is the inward normal.

3.3 Balloon-Type Flow

The similarity-to-template flow that we have just considered has two drawbacks. First, it has no incentive to find the maximal region corresponding to the given template. Second, the flow field does not vanish around the boundary and, in general, the curve will oscillate near the boundary. A possible solution is to add to the objective function in (15) an area term (properly weighted), resulting in an addition of balloon force [8] to the flow. However, the choice of relative weights between the terms is a well-known issue.

We chose a more robust alternative by considering a normal flow where the speed is governed by a histogram difference metric. At each point (x,y) on the curve, a local histogram $\mathrm{Hist}(x,y)$ of feature values is extracted. This may be done efficiently using the integral histogram data structure [26] as will be described in the next section. If H is the template histogram, then the flow at (x,y) is

$$\vec{F}(x,y) = -e^{-\alpha \mathrm{EMD}(\mathrm{Hist}(x,y),H)}\vec{N}. \tag{18}$$

This way, in a sense, we compute the weighted distance from the original contour. Once the distance map has been computed, we would like to extract a contour along which integration over the distance gradient is the largest. Here, we implemented a heuristic approach that goes back to Malladi et al. [22] and Caselles et al. [3], where we actually stop the computation of the distance as the distance gradient gets higher than a specific threshold. In fact, since here we do not have the parabolic diffusion terms in the above models, the treatment of the problem as that of a search for high distance gradients along a distance map becomes natural and provides a new variational meaning to the proposed solution.

We note that the local histograms $\mathrm{Hist}(x,y)$ are extracted from a neighborhood whose size is chosen empirically. The neighborhood should be large enough to provide a meaningful histogram, but not too large to obtain good localization.

4 Initialization

We now present a method for initializing curves for the maximal-discrepancy functional. Using the integral histogram data structure [26], we may efficiently associate with each pixel a local histogram representing the distribution of features over a rectangular patch around the pixel. This histogram serves as an easily computable local feature.

We use this idea to derive the following variant of k-means clustering. First, we cover the image by a set of M square (or rectangular) neighborhoods $\{N(x_i,y_i)|i=1,\ldots,M\}$ with centers at (x_i,y_i) and half widths/heights w. For simplicity, we assume that (x_i,y_i) are spaced $2w+1$ pixels apart, so that every pixel in the image belongs to exactly one neighborhood.

Next, we extract the vector Q_i corresponding to the per-bin counts in the neighborhood $N(x_i,y_i)$. Q_i is of length N (the number of bins in the histograms) and its kth coordinate counts the number of pixels in $N(x_i,y_i)$ falling into bin k. The following is a k-means algorithm for clustering the local histograms $\{Q_1,\ldots,Q_M\}$:

Algorithm 3.

1. Initialize two cluster centers C_1, C_2 by choosing randomly from $\{Q_1,\ldots,Q_M\}$.
2. For each Q_i, compare $\mathrm{EMD}(Q_i,C_1)$ and $\mathrm{EMD}(Q_i,C_2)$. Assign Q_i to the closer cluster center.
3. Update C_1, C_2 by summing the counts of all the assigned Q_is:

$$C_j^{new} = \sum_{\{i|\mathrm{label}(i)=j\}} Q_i. \tag{19}$$

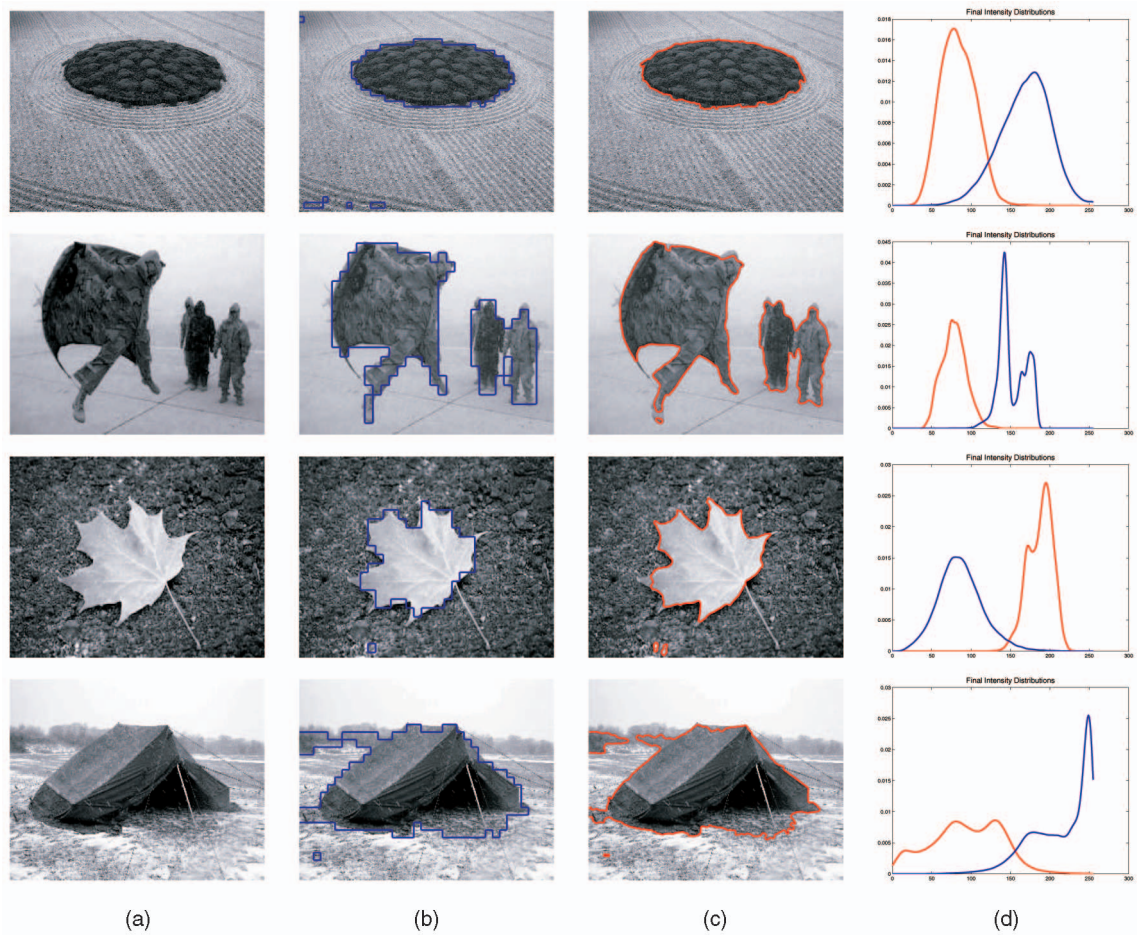

Fig. 1. Maximal-discrepancy functional. (a) Gray-scale images, only intensity was used. (b) Result of initialization stage (Algorithm 3). (c) Results after running the maximal-discrepancy flow (Algorithm 1). (d) Final foreground and background intensity distributions.

Note that here we use the fact that the descriptors Q_i are counts and were not normalized to histograms.

4. Go back to step 2 until assignments have stabilized.

By running this algorithm, we get a partition of our M local neighborhoods into two sets. The quality of this partition is the EMD between the two final vectors C_1 and C_2. We run this algorithm R times—each time with a different random initialization of the two cluster centers—and choose the partition with maximal final EMD between C_1 and C_2.

5 Implementation

The first two gradient flows derived in Section 3 were implemented using an explicit scheme within the level-sets framework [30], [25]. For the local histogram similarity flow, we used the fast marching method, since the curve passes through each point only once.

The choice of time step in the implementation of the EMD flow is restricted by the CFL condition [30], [25], and is limited by the largest flow magnitude in the relevant spatial domain. If the outer and inner regions have significantly different areas, the flow field may exhibit large variations in magnitude, leading to slow evolution. We, therefore, bound the outer region such that $A_{out} \approx A_{in}$.

6 Experimental Results—Still Images

We first experiment with the maximal-discrepancy functional and then demonstrate the match-to-template functional (with balloon force flow) using segmentation in ultrasound images and scene segmentation as discussed in Section 1.

Fig. 1 shows several gray-scale images. We ran the initialization method described in Section 4 (Algorithm 3), where the local feature vector is a 32-bin intensity histogram, extracted on a 9×9 neighborhood ($w = 4$). Column (b) shows the results of this initialization stage, which indeed provide a good starting point for the flow.

In column (c) of Fig. 1, we see the result of employing the maximal-discrepancy flow (Algorithm 1), starting from the curve shown in column (b). We stopped the flow upon convergence of the functional value $\mathrm{EMD}(\Omega)$ (3) to a local maximum. Column (d) shows the kernel-density-estimated intensity histograms for the final foreground and background regions. Recall that the flow is designed to maximize the difference between these two distributions. We emphasize that in all of these results we used only gray-scale intensities quantized to 32 bins, and all parameter values were the same.

Sometimes, the maximal-discrepancy functional does not capture a perceptually meaningful region of the image. Fig. 2 shows two such examples, where the flow separates the dark and bright regions, giving a flawed segmentation.

Fig. 3 demonstrates the advantage that the EMD flow has over traditional flows for bin-wise metrics as the Bhattacharyya flow or Kullback-Leibler flow [16], [24]. We initialized the yellow contour, and ran the EMD, Bhattacharyya, and Kullback-Leibler flows to maximize the distributional difference between foreground and background. When using bin-wise metrics as the Bhattacharyya or Kullback-Leibler metrics, the resulting segmentations in columns (b) and (c) are as good a result (in terms of the functional maximal

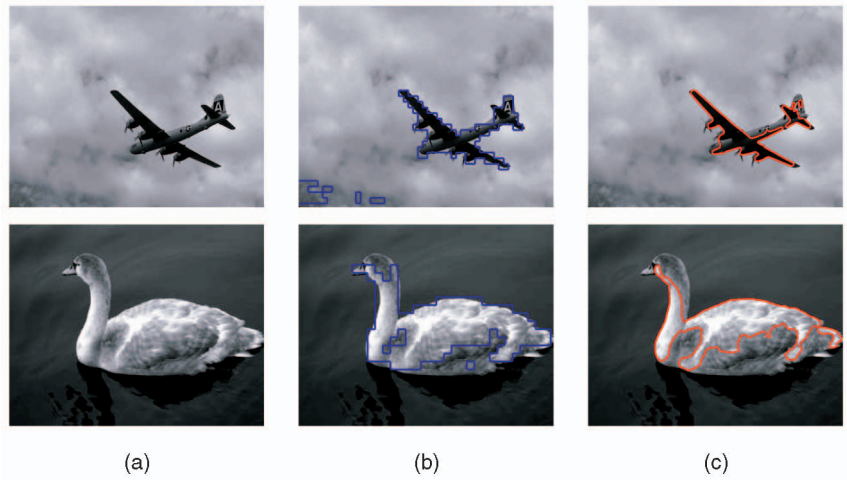

Fig. 2. Examples where the maximal-discrepancy functional has limited perceptual value. Note that the flow separates dark from bright regions thus maximizing the discrepancy between distributions.

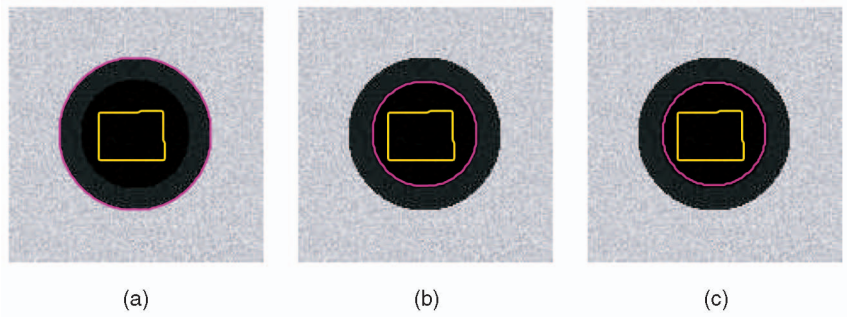

Fig. 3. Initial contour (yellow) and final segmentation for three flows. (a) EMD-based flow. (b) Bhattacharyya flow. (c) Kullback-Leibler flow. Note that EMD flow converged to a solution which perceptually seems to separate foreground from background better than the solution obtained by the flows based on bin-wise metrics.

value) as the segmentation in column (a). However, with the EMD metric, the segmentation in column (a) provides a larger value for the functional than those in columns (b) and (c), and indeed, the EMD flow is the only flow to converge to this segmentation.

We ran the balloon-type flow (Section 3.3) on intracardiac ultrasound images. In this application, segmentation is a first stage before further processing such as 3D reconstruction [2]. The segmentation process is semisupervised in two aspects. First, the initial "seed" is placed by the user inside the region to be segmented. Second, the user chooses the time to stop the flow. This extra input from the user is required since usually parts of the object boundary are nonexistent, and hence, the curve flows beyond the object. Automatic methods for handling this problem—for example, shape-based prior knowledge [28]—are out of the scope of this work. Fig. 4 shows some results. One may note the quality of the results on these challenging images, using 32-bin quantized gray-scale intensities.

7 Activity-Based Scene Segmentation

We now return to the application discussed earlier—namely spatial segmentation of a scene based on a video showing scene activity. We will describe the features whose distributions we compare using the EMD, and hence, derive the flow of the segmenting curve.

7.1 Feature Extraction

Let $\{I(x,y,t)\}$ be a video sequence from a given scene. We assume that the camera is static, as is the case in most surveillance videos (small camera vibrations due to wind may be compensated using standard alignment techniques). The features for which we will compute one-dimensional histograms are normalized values of the spatiotemporal derivatives I_x, I_y, I_t. These features were used by Zelnik-Manor and Irani [31] for action representation and recognition. In contrast to Zelnik-Manor and Irani [31], who used a single global histogram of these features, we use local histograms of these features as local features of a spatial position. In order to efficiently extract these local descriptors over multiple positions, we use the integral histogram data structure as described in Section 4.

The following is a detailed description of the feature extraction process:

1. Input: a video sequence $\{I(x,y,t)|t=1,\ldots,F\}$, number of bins B, temporal change threshold T.
2. Output: integral histograms $\mathrm{IH}_x, \mathrm{IH}_y, \mathrm{IH}_t$ each containing B integral images for each bin $b=1,\ldots,B$.
3. For each current frame $I(\cdot,\cdot,k)$, compute the following three normalized derivative images:

$$N_x = \frac{|I_x|}{\sqrt{{I_x}^2+{I_y}^2+{I_t}^2}}, \quad N_y = \frac{|I_y|}{\sqrt{{I_x}^2+{I_y}^2+{I_t}^2}},$$
$$N_t = \frac{|I_t|}{\sqrt{{I_x}^2+{I_y}^2+{I_t}^2}}.$$

4. Bin every pixel in each of the images N_x, N_y, N_t into one of B bins, taking into account only pixels where $|I_t| > T$.
5. Compute the integral histograms of the current frame I obtaining $\mathrm{IH}_x^k, \mathrm{IH}_y^k, \mathrm{IH}_t^k$.
6. Aggregate the per-frame integral histograms over all frames

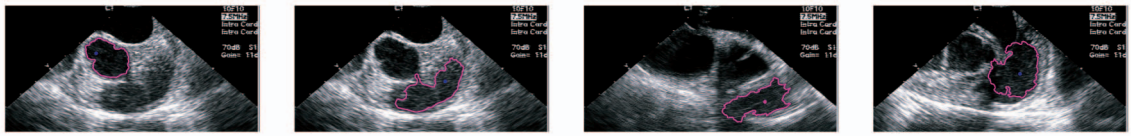

Fig. 4. Intracardiac ultrasound segmentation results. The curve has grown from a user-selected seed (marked) using the EMD similarity-to-template flow (Section 3.3).

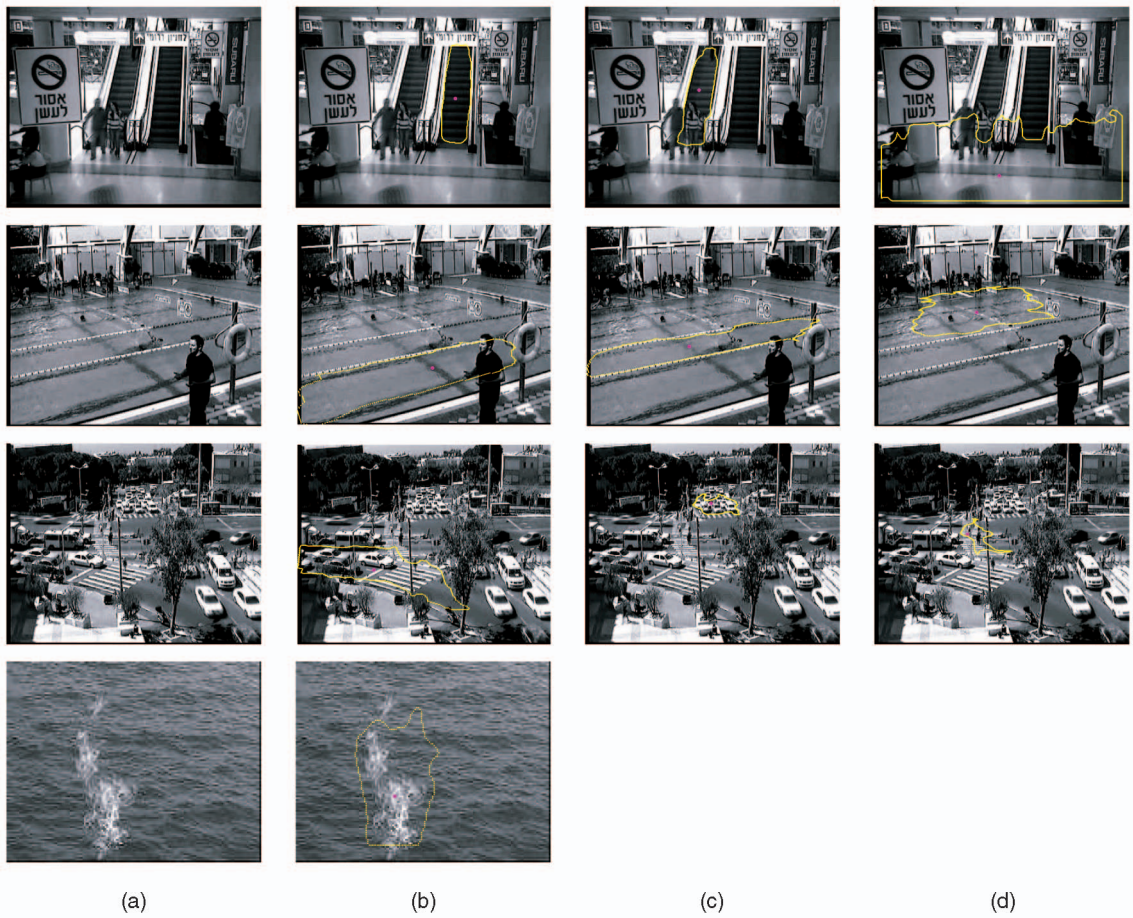

Fig. 5. Example scenes and extracted regions in them. (a) First frame from the input videos. (b)-(d) Example regions. These regions are associated with different activities, and have a clear semantic meaning.

$$1 \le k \le F : \mathrm{IH}_x = \sum_k \mathrm{IH}_x^k, \mathrm{IH}_y = \sum_k \mathrm{IH}_y^k, \mathrm{IH}_t = \sum_k \mathrm{IH}_t^k.$$

Note that each data structure in the sum is actually B images—one for every bin.

The above process (appropriately implemented) can run in real time on a video sequence. The output is three integral image data structures $\mathrm{IH}_x, \mathrm{IH}_y, \mathrm{IH}_t$ allowing us to efficiently extract the histogram of N_x, N_y, or N_t values over a rectangular spatial region extending over all frames $k = 1, \ldots, F$.

Real-life scenes often exhibit large variations in the dynamics of objects in the scene. For example, urban outdoor scenes contain both slow pedestrians and fast vehicles. Therefore, we run the above process over several (usually two or three) temporal scales. This is done by temporal smoothing and downsampling of the sequence.

7.2 Similarity-to-Template Image

Suppose that we extracted $(\mathrm{IH}_x^1, \mathrm{IH}_y^1, \mathrm{IH}_t^1, \ldots, \mathrm{IH}_x^R, \mathrm{IH}_y^R, \mathrm{IH}_t^R)$, where $r = 1, 2, \ldots, R$ denotes the temporal resolution. We have $3R$ integral histograms which allow us to extract $3R$ histograms from each rectangular region. Let $\mathrm{Hist}_i(x, y)$ denote the ith histogram extracted from a rectangular neighborhood of (x, y) (so that $\mathrm{Hist}_1(x, y)$ is the histogram of N_x values at first temporal resolution, $\mathrm{Hist}_6(x, y)$ is the histogram of N_t values at second temporal resolution and so on). Let H_i be the ith template histogram, extracted from a template region selected by the user. Then, the similarity image to this template is

$$S(x, y) = \sum_{i=1}^{3R} \mathrm{EMD}(\mathrm{Hist}_i(x, y), H_i). \tag{20}$$

From this similarity image, we derive the flow

$$\vec{F}(x, y) = -e^{-\alpha S(x,y)} \vec{N}, \tag{21}$$

as in (18).

7.3 Results

We now present several examples of semisupervised spatial segmentation of scenes using video input. Fig. 5 shows several example scenes and regions segmented in these scenes. The first column shows the first frame from the input video. The next columns show the user-selected seed, and the region that was

segmented using that seed. In the first row (mall scene), we segmented two escalators and a floor region where people move freely. These regions clearly have different activities associated with them.

In the second row (pool scene), the pool has three distinct regions. In the far region, children play freely and no swimmers exercise. The second-from-camera lane is intended for fast swimmers and the closest lane is reserved for slow swimmers. Given user-chosen seeds, our algorithm extracted these regions successfully. Note that no overflow to nonwater areas has occurred. Also note that the lane markings are there in the images but were not used and results would have been the same had they not been in the pool.

The third row depicts a junction scene. In this scene, we segmented a right-turning lane where the traffic is generally fast (b), several lanes going downward in the image, where traffic is either fast or standing (corresponding to green/red lights) (c), and a pedestrian-crossing zone (d). Again, these regions have clear semantic meaning and could not have been segmented without using video data.

Finally, we show in the fourth row that our method is successful on classical dynamic textures such as those in [14].

8 Discussion and Summary

Motivated by medical and surveillance applications, we explored in this work the issue of histograms-based curve evolution. Using a tractable form of the EMD for one-dimensional histograms, we derived gradient flows for several functionals using the EMD.

The maximal-discrepancy functional is mathematically appealing and was shown to be perceptually valid in some cases, usually of uniform background. However, we also presented cases where it was shown to be of limited perceptual value. We found the similarity-to-template functional useful in both of the applications we considered, using a variant to maximize the area similar to a template.

In applying the similarity-to-template flow, we used the integral histogram to efficiently extract local histograms all over the image. The same data structure allows efficient initialization of curves as we have demonstrated using a variant of k-means.

The efficient and simple form of EMD, together with the use of the integral histogram, allow the segmentation process to run in real-time, which is extremely important in the medical application we considered (being done as part of a medical intervention).

Finally, in surveillance-type scenes, we demonstrated extraction of regions that have a clear semantic/activity interpretation. We did this using nonparametric local representations using multiple one-dimensional histograms of normalized spatiotemporal derivatives. On real-life challenging scenes, our results compare favorably with state-of-the-art approaches which use representations based on dynamic systems [14], [5].

Acknowledgments

The authors thank Gianfranco Doretto for motivating this work [14] and for providing the fire-over-water sequence. This research was partly supported by the Horowitz fund and by the Israel Ministry of Science research networks program under the Medical and Biological Imaging grant no. 3-3414.

References

[1] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson, "Image Segmentation Using Active Contours: Calculus of Variations or Shape Gradients?" *SIAM J. Applied Math.,* vol. 63, no. 6, pp. 2128-2154, 2003.

[2] E. Brem, BioSense Webster, Inc., personal communication, http://www.biosensewebster.com/products/navigation/cartosound.aspx, 2007.

[3] V. Caselles, F. Catte, T. Coll, and F. Dibos, "A Geometric Model for Active Contours," *Numerische Math.,* vol. 66, pp. 1-31, 1993.

[4] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic Active Contours," *Int'l J. Computer Vision,* vol. 22, no. 1, pp. 61-79, 1997.

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

[6] T. Chan, S. Esedoglu, and K. Ni, "Histogram Based Segmentation Using Wasserstein Distances," *Proc. Conf. Scale Space Methods and Variational Methods in Computer Vision,* 2007.

[7] T. Chan and L. Vese, "Active Contours without Edges," *IEEE Trans. Image Processing,* vol. 10, no. 2, pp. 266-277, Feb. 2001.

[8] L. Cohen, "On Active Contour Models and Balloons," *CVGIP: Image Understanding,* vol. 53, no. 2, pp. 211-218, 1991.

[9] L. Cooper, J. Liu, and K. Huang, "Spatial Segmentation of Temporal Texture Using Mixture Linear Models," *Workshop Dynamical Vision (in conjunction with Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[10] D. Cremers, M. Rousson, and R. Deriche, "A Review of Statistical Approaches to Level Set Segmentation: Integrating Color, Texture, Motion and Shape," *Int'l J. Computer Vision,* vol. 72, no. 2, pp. 195-215, 2007.

[11] D. Cremers and S. Soatto, "Motion Competition: A Variational Approach to Piecewise Parametric Motion Segmentation," *Int'l J. Computer Vision,* vol. 62, no. 3, pp. 249-265, 2005.

[12] D. DeMenthon, "Spatio-Temporal Segmentation of Video by Hierarchical Mean Shift Analysis," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2006.

[13] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto, "Dynamic Textures," *Int'l J. Computer Vision,* vol. 51, no. 2, pp. 91-109, 2003.

[14] G. Doretto, D. Cremers, P. Favaro, and S. Soatto, "Dynamic Texture Segmentation," *Proc. IEEE Int'l Conf. Computer Vision,* pp. 1236-1242, 2003.

[15] C. Fowlkes, S. Belongie, and J. Malik, "Efficient Spatiotemporal Grouping Using the Nystrom Method," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2001.

[16] D. Freedman and T. Zhang, "Active Contours for Tracking Distributions," *IEEE Trans. Image Processing,* vol. 13, no. 4, pp. 518-526, Apr. 2004.

[17] H. Greenspan, J. Goldberger, and A. Mayer, "Probabilistic Space-Time Video Modeling via Piecewise GMM," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 26, no. 3, pp. 384-396, Mar. 2004.

[18] S. Jehan-Besson, M. Barlaud, and G. Aubert, "Video Object Segmentation Using Eulerian Region-Based Active Contours," *Proc. IEEE Int'l Conf. Computer Vision,* 2001.

[19] S. Jehan-Besson, M. Barlaud, and G. Aubert, "Dream^{2}s: Deformable Regions Driven by an Eulerian Accurate Minimization Method for Image and Video Segmentation," *Int'l J. Computer Vision,* vol. 53, no. 1, pp. 45-70, 2003.

[20] S. Kahn and M. Shah, "Object Based Segmentation of Video Using Color, Motion and Spatial Information," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2001.

[21] Z. Lu, W. Xie, J. Pei, and J. Huang, "Dynamic Texture Recognition by Spatio-Temporal Multiresolution Histograms," *Proc. IEEE Workshop Applications of Computer Vision,* pp. 241-246, 2005.

[22] R. Malladi, J. Sethian, and B. Vemuri, "Shape Modeling with Front Propagation: A Level Set Approach," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 17, no. 2, pp. 158-175, Feb. 1995.

[23] R. Megret and D. DeMenthon, "A Survey of Spatio-Temporal Grouping Techniques," Technical Report LAMP 094, LAMP—Univ. of Maryland, Aug. 2002.

[24] O. Michaelovich, Y. Rathi, and A. Tannenbaum, "Image Segmentation Using Active Contours Driven by the Bhattacharyya Gradient Flow," *IEEE Trans. Image Processing,* vol. 16, no. 11, pp. 2787-2801, Nov. 2007.

[25] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces.* Springer-Verlag, 2002.

[26] F. Porikli, "Integral Histogram: A Fast Way to Extract Histograms in Cartesian Spaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2005.

[27] R. Ronfard, "Region-Based Strategies for Active Contour Models," *Int'l J. Computer Vision,* vol. 13, no. 2, pp. 229-251, 1994.

[28] M. Rousson and D. Cremers, "Efficient Kernel Density Estimation of Shape and Intensity Priors for Level Set Segmentation," *Proc. Conf. Medical Image Computing and Computer Assisted Intervention,* pp. 757-764, 2005.

[29] Y. Rubner, C. Tomasi, and L. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int'l J. Computer Vision,* vol. 40, no. 2, pp. 91-121, 2000.

[30] J. Sethian, *Level Sets Methods and Fats Marching Methods: Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision and Materials Science,* second ed. Cambridge Univ. Press, 1999.

[31] L. Zelnik-Manor and M. Irani, "Statistical Analysis of Dynamic Actions," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 9, pp. 1530-1535, Sept. 2006.

[32] Y. Zhai and M. Shah, "A General Framework for Temporal Video Scene Segmentation," *Proc. IEEE Int'l Conf. Computer Vision,* 2005.

[33] S.-C. Zhu and A. Yuille, "Region Competition: Unifying Snakes, Region Growing and Bayes/MDL for Multiband Image Segmentation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 9, pp. 884-900, Sept. 1996.