A new fusion framework for motion segmentation in dynamic scenes

Lazhar Khelifi, and Max Mignotte

***Abstract*—Dynamic texture (DT) segmentation, and video processing in general, is currently widely dominated by methods based on deep neural networks with a large number of layers. Although this parametric approach has achieved great success on the dynamic texture segmentation, all current deep learning methods suffer from a significant main weakness related to the lack of a sufficient reference annotation to train models and to make them functional. Annotation task is time-consuming and tedious. In particular, it requires highly experienced and professional work to build a large dataset of images specifically annotated for each object type or class. In addition, the result of these methods can deteriorate significantly when the network is fed with images or video not similar (in terms of color, shape, texture, etc.) to the images previously included in the training set. This paper explores the unsupervised segmentation approach that can be used in the absence of training data to segment new videos. In particular, it tackles the task of dynamic texture segmentation : clustering into groups various characteristics and phenomena that reproduce in both time and space, assigning a unique label to each group or region. We present an effective unsupervised learning consensus model for dynamic texture segmentation (ULCM), whose aim is to fuse multiple and weak region-based segmentation maps to get a final better segmentation result. The different label fields to be combined, are given by a simple clustering technique applied to an input video (based on three orthogonal planes xy, xt and yt). The proposed model uses as features the set of values of the requantized local binary patterns (LBP) histogram around the pixel to be classified. We perform experiments on the challenging SynthDB dataset which show that ULCM is faster, easy to implement, simple and has few parameters, compared to existing dynamic texture segmentation approaches that require that require either a parameter estimation or a training step. In addition, qualitative experiments on the YUP++ dataset show that ULCM obtains competitive results.**

***Index Terms*—Video processing, dynamic texture segmentation, consensus framework, unsupervised learning, optimization, global consistency error (GCE).**

I. Introduction

Dynamic texture (or texture movie) combines texture in the spatial domain with motion (with some form of stationarity) in the temporal domain [1] (see Fig. 1). Consequently, dynamic texture segmentation can be very complex because this process requires to jointly analyze spatiotemporal data which can be very different in nature, just like the numerous dynamic scenes existing in the real world, such as; cloud, falling snow, flowing flag, swirl, smoke, etc. [2]. Recently, research on dynamic segmentation of textures has been of growing interest and has led to the development of interesting and varied methods. Doretto *et al.* [3] used spatio-temporal statistics and more precisely their dynamics over time with Gauss-Markov models to segment a sequence of images into regions. A variational optimization framework was then used to infer the parameters of the model and to localize the boundary of each region. However, a limitation of this model is based on the assumption that regions vary slowly over time and also essentially according to the irradiance in each region. Vidal *et al.* [4], for their part, tackled this problem by first analyzing a generalized principal component analysis (GPCA) of the optical flow field of the video which was finally exploited to segment the spatiotemporal data by grouping pixels having similar trajectories in time. Nevertheless, as it was originally designed, this segmentation model is limited to only two classes. Chan *et al.* [5] proposed the mixture of dynamic textures (DTM) as a suitable representation for both the appearance and dynamics of dynamic texture videos. They used an expectation-maximization (EM) algorithm for learning the parameters of the model. Their work was extended in [6] by using the efficiency of the GPU computations to accelerate the segmentation process. Wattanachote *et al.* [7] presented a new and original semiautomatic dynamic texture segmentation method by exploiting motion vectors derived from Farnebäck's [1] model [8]. Nevertheless, an important limitation of this technique is that the intervention of the user remains necessary to select the target objects and to adjust the result to produce a high quality spatiotemporal segmentation map. Nguyen *et al.* [9] proposed a new unsupervised feature selection dynamic mixture model (FSDTM) for motion segmentation. The main advantage of their method is that it is totally unsupervised and does not require a set of training data having known classifications on which to fit the mixture model. In this approach, the Expectation Maximization (EM) algorithm is exploited to estimate the parameters of the mixture model in the maximum likelihood sense. However, the EM algorithm remains very sensitive to initial values, noise, outliers and to the shapes of the laws of distribution chosen a priori in the mixture model and has also the drawback of converging at local minima. An interesting (but partially supervised) approach combining a filter-based motion features with a supervised learning approach has been introduced by Teney *et al.* [10]. Different from the existing methods, Cai *et al.* [11] have proposed a new dynamic texture method for ultrasound images. This model is based on surfacelet transform, HMT model and parallel

The authors are with the Department of Computer Science and Operations Research, Faculty of Arts and Sciences, University of Montreal, Montreal, QC H3C 3J7, Canada (e-mail: khelifil@iro.umontreal.ca; mignotte@iro.umontreal.ca)

[1] An algorithm for estimating dense optical flow based on modeling the neighborhoods of each pixel by quadratic polynomials.

computing. One advantage of this approach is that it makes it possible to use both spatial and temporal information of coefficients into one model by considering simultaneously a sequence of images. Yousefi *et al.* [12] proposed a novel non-parametric fully Bayesian approach for DT segmentation, formulated on the basis of a joint Dirichlet process mixture (DPM) with generative dynamic texture models (GDTMs). This method eliminates efficiently the expert knowledge about the number of the dynamic textures and initial partitioning. In [13] authors discussed three DT segmentation methods based on optical flow, local spatiotemporal technique (local binary pattern) and global spatiotemporal technique (Contourlet transform). Their experimentation is carried out using these individual techniques and also with some combinations. Results showed that optical flow technique is computationally more complex but a natural way of detecting motion. Contrary, local binary pattern is computationally less complex and simple to implement and a suitable variant can be considered depending on the application at hand. This study also showed that Contourlet Transform works well with Natural DTs as it has the capability of tracing smooth contours in the image. Among the most recent work, one can cite the algorithm proposed by Andrearczyk *et al.* [14] in which a CNNs is applied on three orthogonal planes xy, xt and yt of the video sequence. The major drawback in their approach is that the training of independent CNNs on three orthogonal planes, and the combination of their outputs makes the process more complex from a computational point of view while being also supervised. Motivated by the above observations, we herein introduce a new fusion model for dynamic texture segmentation called ULCM. Our model aims to combine multiple and weak segmentation results in order to obtain a more reliable and high-quality spatiotemporal segmentation map. These initial and weak segmentation results are estimated from different frames (or slices) of the video sequence and across the different axis of the data cube. In addition, in order to overcome the disadvantages of previous methods that often lead to complex estimation, optimization or combinatorial problems, we herein propose a simple energy-based model based on an efficient segmentation fusion criterion derived from the Global Consistency Error (GCE). The GCE criterion is a perceptual measure which takes into account the inherent multiscale nature of any image segmentation (which could be possibly viewed as a refinement of another segmentation) by measuring the level of difference between two segmentation maps. In addition, to efficiently optimize our energy-based model, we propose a modified local optimization procedure derived from the iterative conditional modes (ICM) algorithm.

In summary, this paper makes the following main contributions:

- We propose a new unsupervised learning consensus model for dynamic texture segmentation. Our model aims to combine multiple and weak segmentation results to achieve a more reliable and final refined segmentation of an input video.
- We use an energy function originated from the global consistency error (GCE). The GCE criterion is a perceptual measure which takes into account the inherent multiscale nature of an image segmentation (by measuring the level of refinement existing between two spatial partitions).
- We evaluate the proposed method over two benchmark datasets. Comprehensive experimental results demonstrate that the proposed method is able to produce high quality segmentation results with clear boundaries and significantly outperforms state-of-the-art approaches.

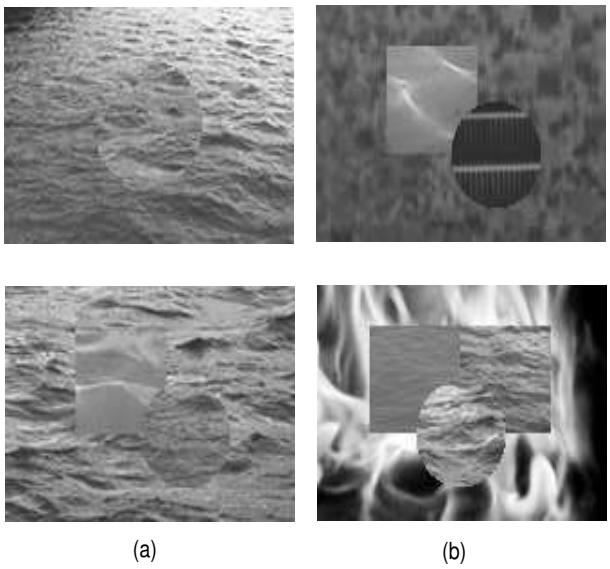

Fig. 1. Examples of DTs. (a) DTs are different in terms of temporal mode, i.e., movement or motion) but similar to their spatial mode (i.e., appearance) related essentially to the texture. (b) DTs are different in terms of spatial mode, but similar to their temporal mode.

The remainder of the paper is organized as follows: We begin with a brief definition of dynamic texture in section II. In Section III, we introduce the ULCM model. In Section IV, we present an experimental evaluation of the proposed algorithm using synthetic and real video datasets. In Section V, we conclude the paper.

II. Dynamic Texture

While a variety of definitions of the dynamic texture have been suggested, this paper will use the definition first suggested by Chan *et al.* [5] who define it as a generative model for both the appearance (video frame at time t), and the dynamics of video sequences (evolution of the video over time), based on a linear dynamic system. While the appearance of frame $y_t \in R^n$ is a linear function of the current state vector, plus some observation noise, the dynamics are represented as a time-evolving state process $x_t \in R^n$ (typically $n \ll m$). Mathematically, the equations of this system are defined as follows:

$$s(x) = \begin{cases} x_{t+1} = Ax_t + v_t \\ y_t = Cx_t + w_t \end{cases} \quad (1)$$

Where, the next value of the state variable x_{t+1} depends on the present value x_t, and the present value of the observation process y_t also depends on x_t. The parameter $A \in R^n$ is a state-transition matrix, and $C \in R^m$ is a matrix containing

the principal components of the video sequence. The driving noise process v_t is normally distributed with zero mean and covariance Q, that is, $v_t \sim N(0; Q)$, where $Q \in R^{n\times n}$ is a positive-definite $n \times n$ matrix. The observation noise w_t is also zero mean and Gaussian, with covariance R, that is, $w_t \sim N(0; R)$, where $R \in R^{m\times m}$. It should be noted that each coordinate of the state vector x_t defines a one-dimensional random trajectory in time. A pixel is then represented as a weighted sum of random trajectories, where the weighting coefficients are contained in the corresponding row of C. The dynamic texture is completely represented as a graphical model in Fig. 2.

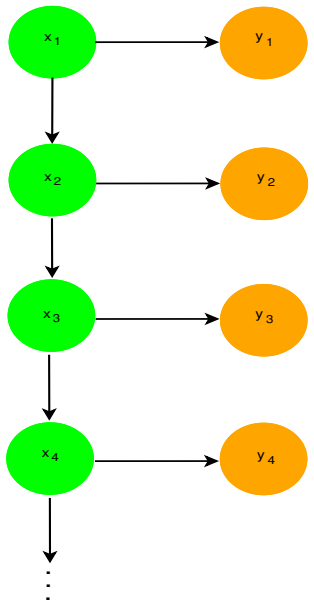

Fig. 2. Graphical model for dynamic texture DT. x_t and y_t are the hidden state and observed video frame at time t

III. Proposed Method

The method described here is automatic, simple, and performed through five steps, as mentioned in our preliminary work [15]. In the first step of our method, a set of images is generated by slicing through the video cube (i.e. dynamic texture data). In the second step, a feature extraction process is proposed and performed for each image. In the third step, a different stochastic dimensionality reduction based on different seeds is applied to the extracted local histogram associated with each pixel. Then, a set of initial segmentations is generated by a clustering technique. As soon as these steps have been carried out, in the fourth step, an energy-based fusion scheme is performed through the set of segmentation maps by iteratively optimizing a deterministic gradient-based optimization algorithm. The pseudo-code of our method is outlined in Fig. 3.

A. Slicing the Dynamic Texture Data

In order to fully benefit from the complementarity of the three intrinsic (spatial and temporal) dimensions of our input video sequence V, and thus to more effectively represent each dynamic texture, we perform the following simple slicing operation: In addition to the classical slicing; in which in the xy spatial plane, we simply generate w equidistant slices equally spaced in the t time axis from V corresponding to the w images contained in the video sequence, we have added two more clipping processes: First, in the time plane xt, we generate h equidistant slices (or frames) equally spaced on the y axis. By this fact, a slice of the xt plane represents the evolution of a line of pixels over time along the video. Second, in the time plane yt, we generate m equidistant slices equally spaced on the x axis. Concretely, a slice of the yt plane represents the evolution of a column of pixels over time along the video sequence. Finally, after this slicing step, we get $h\ timesw\ timesm$ separate images in three sets (see Fig. 2).

B. LBP Representation

To more effectively describe the texture, we apply the local binary pattern (LBP) operator to each previously generated frame (see Fig. 4.(e)). The purpose of the LBP operator is to represent the statistics of the micro-patterns contained in an image (that is, a frame in our case) by encoding the difference between the pixel value of the center point and that of its neighbors [16]. Let F a gray frame and q_c be the value of the center pixel c of a local neighborhood. Let q_p $(p = 0, ..., P-1)$ be the values of P equidistant pixels uniformly distributed around a circle with radius R forming a circularly symmetric set of neighbors. If the coordinates of q_c are $(0, 0)$, then the coordinates of q_p are defined by $(R\sin(\frac{2\pi p}{P}), R\cos(\frac{2\pi p}{P}))$ and the values of neighbors that do not fall exactly on pixels are estimated by bilinear interpolation. The LBP descriptor on this pixel (c) is defined by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(q_s - q_c)2^p,\ s(x) = \begin{cases} 1 & ,x \geq 0 \\ 0 & ,x < 0 \end{cases} \quad (2)$$

C. Generation of the Segmentation Ensemble

Once the LBP representation step is achieved, we project all pixels of each LBP-frame onto the xy plane (cf. Fig. 3.(d)). Then, for each frame and for each pixel, we estimate, within an overlapping squared fixed-size ($Nw = 7$ neighborhood centered around the pixel to be classified, a local requantized LBP histogram. In the next step, we concatenate all local histograms related to the same pixel $p^i_{(x,y)}$, at each time t to finally form a high-dimensional feature vector or a histogram (cf. Fig. 3.(e)). This high-dimensional histogram encloses a wide range of redundant features (or information) and hide the correlations between data which can make the interpretation of data much harder. For that reason, the dimensionality reduction methods can be typically utilized here to avoid the lack of discrimination, often refereed to the so-called "*curse of dimensionality* problem" [3]. In light of this situation, the original features should be preprocessed to simplify the high-dimensional histograms by finding low-dimensional structure with it. Also, it should be noted that, the precision of the segmentation must remain satisfactory while reducing the

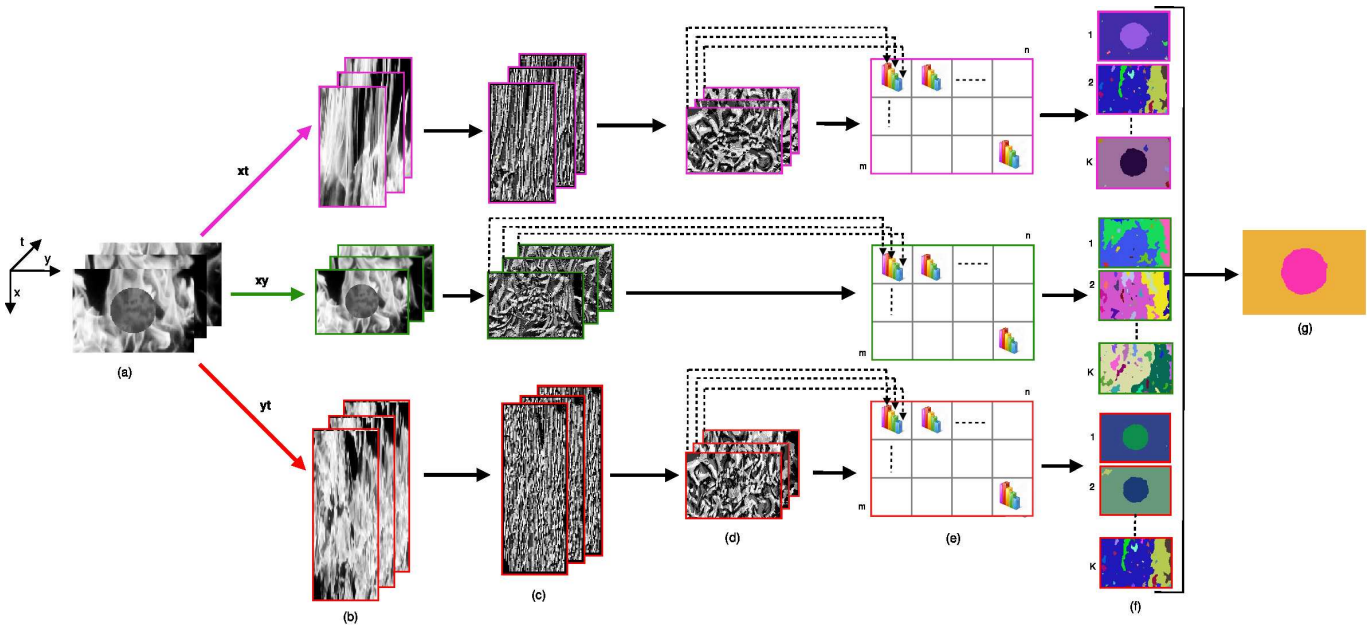

Fig. 3. Proposed system overview. (a) Input video. (b) Slicing step. (c) LBP representation. (d) Projection of LBP frames on the xy plan. (e) Feature extraction and dimensionality reduction.(f) Clustering with k-means. (g) Final combined result.

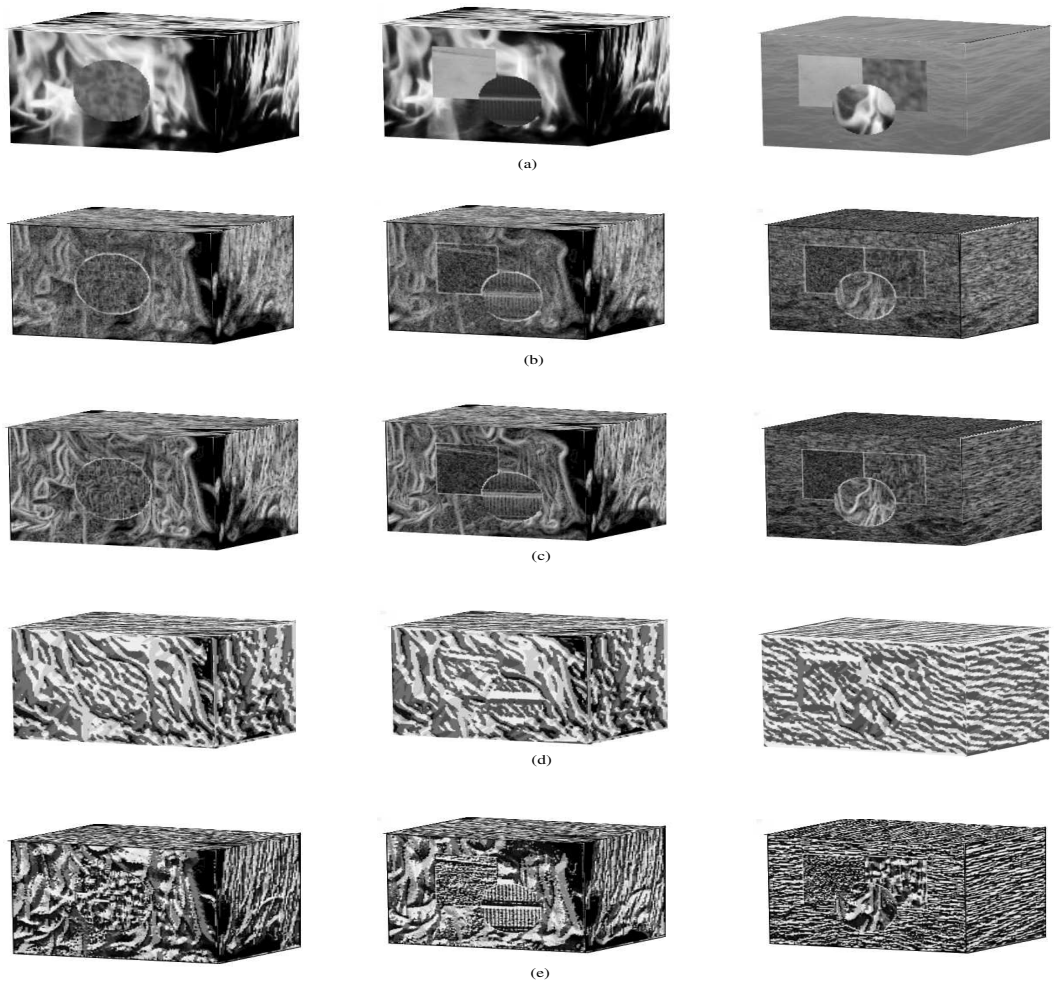

Fig. 4. Representation of the input video with different texture operator. (a) Original video, (b) Histogram of oriented gradients HOG, (c) Laplacian operator LAP, (d) local phase quantization LPQ, (e) local binary pattern LBP.

amount of features to be processed, and eliminating some redundant information. In fact, dimensionality reduction is simply the process of projecting the n-dimensional data onto a subspace of a considerably less lower dimension (k) that represent a set of principal variables. The commonly used approaches include principal component analysis (PCA) [17], multidimensional scaling (MDS) [18] and random projection (RP) [20] [19]. In our work we resort to the random projection (RP) for dimensionality reduction for two reasons. Firstly, RP is a much faster and less complex (linear complexity) compared to MDS and PCA (quadratic complexity). Secondly, RP has the ability to generate, by using different seeds, different (and low-dimensional) noisy projected data which will provide the necessary variability to our algorithm which then use it efficiently to obtain a final robust segmentation (this will be explicit in the following). Mathematically, in the random projection process, the original data matrix X $[n \times m]$ is multiplied by a random projection matrix RP $[m \times k]$ as follows:

$$X_{red} = \frac{1}{\sqrt{k}} \times X \times RP \tag{3}$$

where X_{red} is the result of the projection of the data onto a lower k dimensional subspace. Once the dimension reduction step is done, we pass the various low-dimensional histograms[2](related to the different seeds) to the clustering algorithm to generate groups. At this point, we resort to the useful k-means-based clustering technique [22]. We have adopted this choice to ensure a reduced computational time and cost for this important step.

D. Fusion Based on the Global Consistency Error Criterion

Once the segmentation set has been generated, we undertake to merge or combine all these weak segmentations in an energy-based fusion model under the global consistency error (GCE) criterion.

1) Global Consistency Error Criterion: The GCE criterion is initially derived from the so-called local refinement error (LRE) which tries to quantify the degree of similarity, in term of refinement, between two segmentations [23]. According to this perceptual criterion, segmentations are considered to be consistent when they represent the same segmented image at different levels of detail (or scale) [24] [25] or in other words when they represent a more or less detailed version of the same segmentation. Denote as n the number of pixels within the frame F and let $\Phi_\mu = \{s_\mu^1, s_\mu^2, \ldots, s_\mu^{nb_\mu}\}$ & $\Phi_\nu = \{s_\nu^1, s_\nu^2, \ldots, s_\nu^{nb_\nu}\}$ be, two segmentation results of the same frame to be compared, nb_μ being the number of segments in Φ_μ and nb_ν the number of segments in Φ_ν. Let now p_i be a particular pixel and the couple ($s_\mu^{<p_i>}$,$s_\nu^{<p_i>}$) be the two segments including this pixel, respectively in Φ_μ and Φ_ν. The LRE on this pixel p_i is the defined as follows:

$$\text{LRE}(s_\mu, s_\nu, p_i) = \frac{|s_\mu^{<p_i>} \setminus s_\nu^{<p_i>}|}{|s_\mu^{<p_i>}|} \tag{4}$$

where $|X|$ denotes the cardinality of the set of pixels X and $\setminus$ represents the algebraic operator of difference. Particularly, a value of 1 means that the two regions overlap, in an inconsistent manner, on the contrary, an error of 0 expresses that the pixel is practically included in the refinement area [26]. A good way to force all local refinements to go in the same direction is to make the LRE metric symmetric. In doing so, every LRE must be measured at least twice, once in each sense, and this simple strategy finally leads us to the so-called global coherence error (GCE):

$$\text{GCE}^\star(\Phi_\mu, \Phi_\nu) = \frac{1}{2n}\left\{\sum_{i=1}^{n} \text{LRE}(s_\mu, s_\nu, p_i) + \sum_{i=1}^{n} \text{LRE}(s_\nu, s_\mu, p_i)\right\} \tag{5}$$

The GCE* value lies in the range $[0, 1]$. A distance of 0 indicates a high similarity (in terms of level of details) between the two segmentation maps Φ_μ . Φ_ν. While a distance of 1 expresses a poor consistency or correspondence between the two segmentation maps to be compared.

2) Fusion: Let us assume now that $\{\Phi_k\}_{k \leq J} = \{\Phi_1, \Phi_2, \ldots, \Phi_J\}$ represents the ensemble of J different (weak) segmentations to be combined or fused (according to the GCE criterion). Let us recall that $J = 3K$, with K being the number of segmentation maps generated from each set of frames (cf Fig. 3.(f)). As already said, our goal is to get an improved segmentation result $\hat{\Phi}$ for the video sequence V. As already stated, our ultimate goal is to get the best possible segmentation map of the video sequence from this set of multiple low-cost and weak segmentations. To estimate this refined segmentation result which in fact represents a consensus or a compromise between these multiple weak segmentations, an original and efficient energy-based model framework is now proposed to allow us to reconcile (or fuse) these segmentations. This model aims to generate a segmentation map solution as close as possible, in terms of the considered GCE*-distance to all the other segmentations $\{\Phi_k\}_{k \leq J}$. In this energy-based framework, if Θ_n designates the set of all possible segmentations using n pixels, the consensus segmentation $\hat{S}_{\overline{\text{GCE}}^\star}$ which is optimal according to the GCE* criterion) is then directly defined as the minimizer of the following cost function $\overline{\text{GCE}}^\star$:

$$\hat{\Phi}_{\overline{\text{GCE}}^\star} = \arg\min_{\Phi \in \Theta_n} \overline{\text{GCE}}^\star\big(\Phi, \{\Phi_k\}_{k \leq J}\big) \tag{6}$$

Our fusion model is thus formulated as an optimization problem involving a highly nonlinear cost function. To optimize this nonlinear function [see Eq (6)], stochastic optimization approaches, such as the simulated annealing [28], the genetic algorithm [36] or the exploration/selection/estimation (ESE) procedure [27] can be efficiently used. These algorithms are guaranteed to find the optimal solution, but with the

[2]The size of the final feature vector is 20 times smaller than the size of the original high-dimensional vector.

[3]The curse of dimensionality is a phenomenon that arises when analyzing data in high-dimensional spaces. Adding dimensions stretches the points apart, making high-dimensional data extremely sparse and uniformly distributed [21]. This sparsity is problematic for any algorithm that requires statistical significance. It is important to note that, organizing and searching data often relies on detecting areas where objects form groups with similar properties (i.e., similar pixels in our case); in high-dimensional data, however, all objects appear to be sparse and dissimilar in different ways.

disadvantage of a huge computing time. Another alternative we have adopted in this work is a deterministic optimization strategy based on the iterative conditional mode (ICM) method proposed by Besag [29] (which is actually also equivalent to a Gauss-Seidel based relaxation scheme), where each pixel's label are updated one at a time [30] [31] In our case, this algorithm has the advantage of being simple to implement while also being fast and efficient in terms of convergence.

IV. Experiments and Discussions

A. Experimental Setup

Evaluation Datasets. We have evaluated our model quantitatively on SynthDB, a synthetic video texture database[4] [5] containing 299 8-bit graylevel videos (image size is 160 $\times$ 110 $\times$ 60 pixels). Video sequences are split into three groups (99 videos with 2 labels, 100 videos with 3 labels, and 100 videos with 4 labels), and a common ground truth template is available for each group. This dataset is very challenging, first because videos are grayscale, and also by the fact that textures exhibit very similar static appearance. In addition, we evaluate qualitatively the proposed method on the YUP++ [32] database.

Evaluation Metric. We also rely on the probabilistic Rand (PR) index [33] for the evaluation of segmentation performance. This metric is widely used in the study of the performance of image (sequence) segmentation algorithms. More precisely, the PR index metric counts the fraction of pairs of pixels whose labeling is identical between two image segmentations to be measured. Mathematically, consider two valid label assignments, an automatic segmentation S_{aut} and a manual segmentation (i.e., ground truth) S_{gt} of N pixels $P = p_1, p_2, ...p_i, ..., p_N$ that assign labels b_i and $b_i^{'}$ respectively to pixel p_i . The Rand index R can be given as the ratio of the number of pairs of pixels having a consistent label relationship in S_{aut} and S_{gt}. Therefore, we can consider the probabilistic rand (PR) index as follows:

$$R(S_{aut}, S_{gt}) = \frac{1}{C_N^2} \sum_{i,j;i<j}^{n} [I(b_i = b_i^{'} \wedge b_j = b_j^{'}) + I(b_i \neq b_i^{'} \wedge b_j \neq b_j^{'})] \quad (7)$$

where I is the identity function, and C_N^2 is the number of possible unique pairs among N data points. A score of one indicates a good result, otherwise, a score of zero indicates a bad segmentation.

B. Discussions

Table I shows that the result achieved by our unsupervised method outperforms the other current state-of-the-art methods, although our method has the advantage of not requiring any supervision and/or specific initialization step. As a result, we obtain an interesting PR score equals to 0.953. Additionally, to qualitatively compare the performance of the proposed method against another set of methods, we present one experiment in Fig. 5. In this experiment, our method is compared to the layered dynamic textures (LDT) [35], the dynamic texture model (DTM) [5], the unsupervised and supervised (based learning metric) approaches proposed in [10]. The result of the proposed method, as shown in the sixth column, is clearly better than that of other methods. In Fig. 6 we present additional segmentation results obtained from the SynthDB dataset based on our proposed method. Results on the complete dataset are available publicly on-line in the website of the corresponding author at the following http address: http://www-etud.iro.umontreal.ca/∼khelifil/ResearchMaterial/consensus-video-seg.html. We have also tested the effects of using different fusion criteria. In Table II, we report the performances yielded by our algorithm based on the GCE, VoI, PRI and the F-measure criteria. This test shows that the GCE is the most reliable criteria that yielding the best PR index. In contrast, the lower PR index is achieved based on the PRI criterion with values equal to 0.911 , 0.743 and 0.710, respectively, for videos with two, three and four labels. As another evaluation test, in Fig. 7 we show different segmentation results of three different video obtained based on these criteria. In fact, compared to the PRI, the VOI and the F-measure based results, the GCE criterion (in (e)) achieves a better qualitative results. This shows clearly that our choice of using this criterion is effective. In addition, in table III, we present the performance of the proposed method using different texture features. As we can see, OLBP and ELBP operator histogram are the features that provide the best PR index scores. In order to test the robustness of the proposed technique against the variability of dynamic objects, we experiment it on the YUP++ database. Thus, in Fig. 8 we present different segmentation results for scenes with Waving Flags, Waterfall and Escalator. Finally, in Fig.10 and n Fig.11 we present a plot of the average PR obtained for each class label (of the SynthDB) and the computing time as a function of the dimension of the histogram of features (k).

In summary, our method has the merit of being simple in terms of implementation and numerical computation, totally unsupervised while being efficient compared to others complex, computationally demanding video segmentation models existing in the literature. In addition, our model remains widely perfectible; either by adding other weak segmentations (to be combined) using other interesting (and possibly complementary) features or by using a more efficient fusion criterion or distance in our energy-based fusion framework.

V. Summary and Conclusions

We have presented a new approach to segment video with dynamic textures. By combining multiple and weak region-based segmentation maps of a video, we demonstrated that it is possible to get a final better segmentation result using a new geometric criterion. Experiments show that our model, while being simple, fully unsupervised, fast and perfectible, is comparable to the state of the art methods using supervised or semi-automatic strategy and even better than those relying on unsupervised approaches. A possible extension of this work is

[4]The synthetic video texture database is publicly accessible via this link: http://www.svcl.ucsd.edu/projects/motiondytex/

TABLE I
COMPARISON OF THE PROPOSED METHOD WITH OTHER METHODS ON THE SYNTHDB DATASET (PR INDEX, HIGHER IS BETTER).

ALGORITHMS	PERFORMANCE (Avg. PR)		
	99 videos 2 labels	100 videos 3 labels	100 videos 4 labels
GPCA [4] in [34]	0.515	0.477	0.526
DTM [5]	0.907	0.847	0.859
Color (Unsupervised) [10]	N/A	0.599	N/A
Color + motion (Unsupervised) [10]	N/A	0.727	N/A
Color + motion (Learned, logistic regression) [10]	N/A	0.771	N/A
Color+mouvment (Unsupervised) [10]	0.7113	0.608	0.612
Color+HoME+mouvment (Unsupervised) [10]	0.863	0.795	0.744
-**Proposed method**-	0.953	0.855	0.796

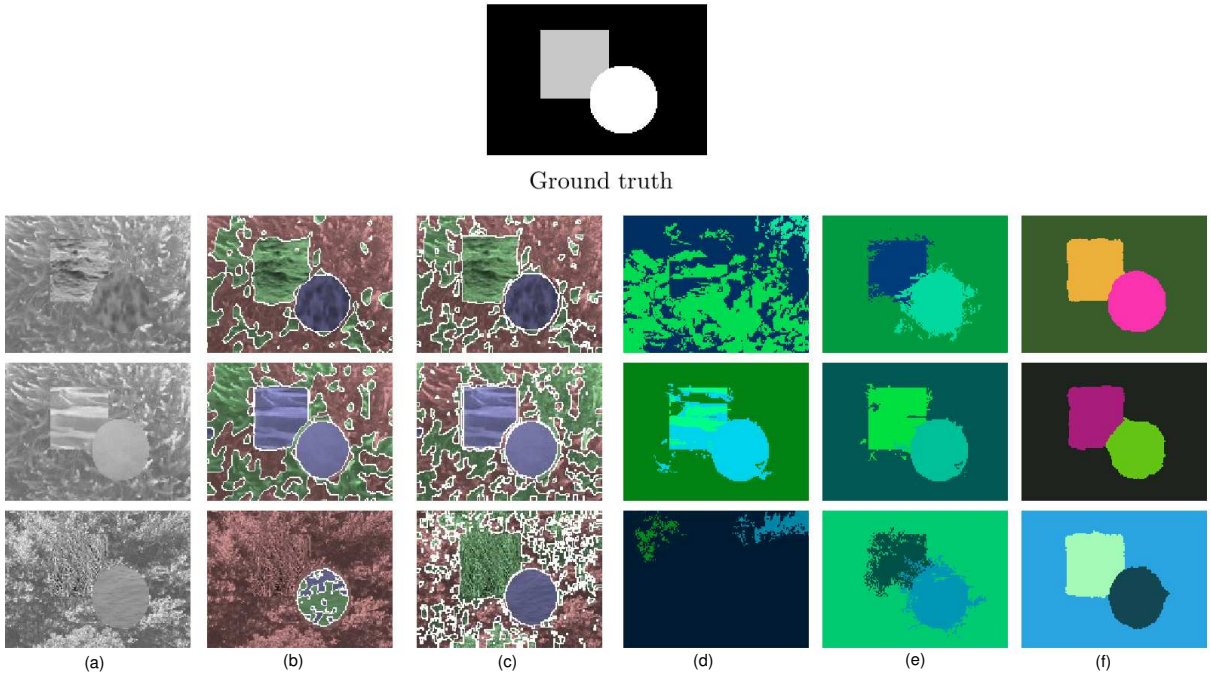

Fig. 5. Examples of segmentation results obtained by our proposed method of three videos (with 3 labels) from the SynthDB dataset [5] compared to other algorithms. (a) Input video, (b) LDT with manual initialization [35], (c) DTM with contour initialization [5], (d) Color+motion Unsupervised [10] (f), Color+motion Learned [10], (e) Proposed method Unsupervised.

TABLE II
PERFORMANCE OF THE PROPOSED METHOD USING DIFFERENT FUSION CRITERIA ON THE SYNTHDB DATASET (PR INDEX, HIGHER IS BETTER).

ALGORITHMS	PERFORMANCE (Avg. PR)		
	99 videos 2 labels	100 videos 3 labels	100 videos 4 labels
-F-measure-	0.937	0.756	0.710
-VoI-	0.947	0.823	0.763
-PRI-	0.919	0.743	0.710
-GCE-	**0.953**	**0.855**	**0.796**

to adopt different types of features with the local binary pattern (LBP) to more represent the dynamic texture. Another possible extension of this work is to combine other possible criteria (variation of information, F-measure and probabilistic rand index) to achieve a more reliable result. It is very important to note that the proposed model is suitable to be implemented in parallel or to fully take advantage of GPU systems that allows simultaneously handling of different types of features or criteria.

REFERENCES

[1] F. Hajati, M. Tavakolian, S. Gheisari, Y. Gao, and A. S. Mian , "Dynamic Texture Comparison Using Derivative Sparse Representation: Application to Video-Based Face Recognition," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 970–982, 2017.

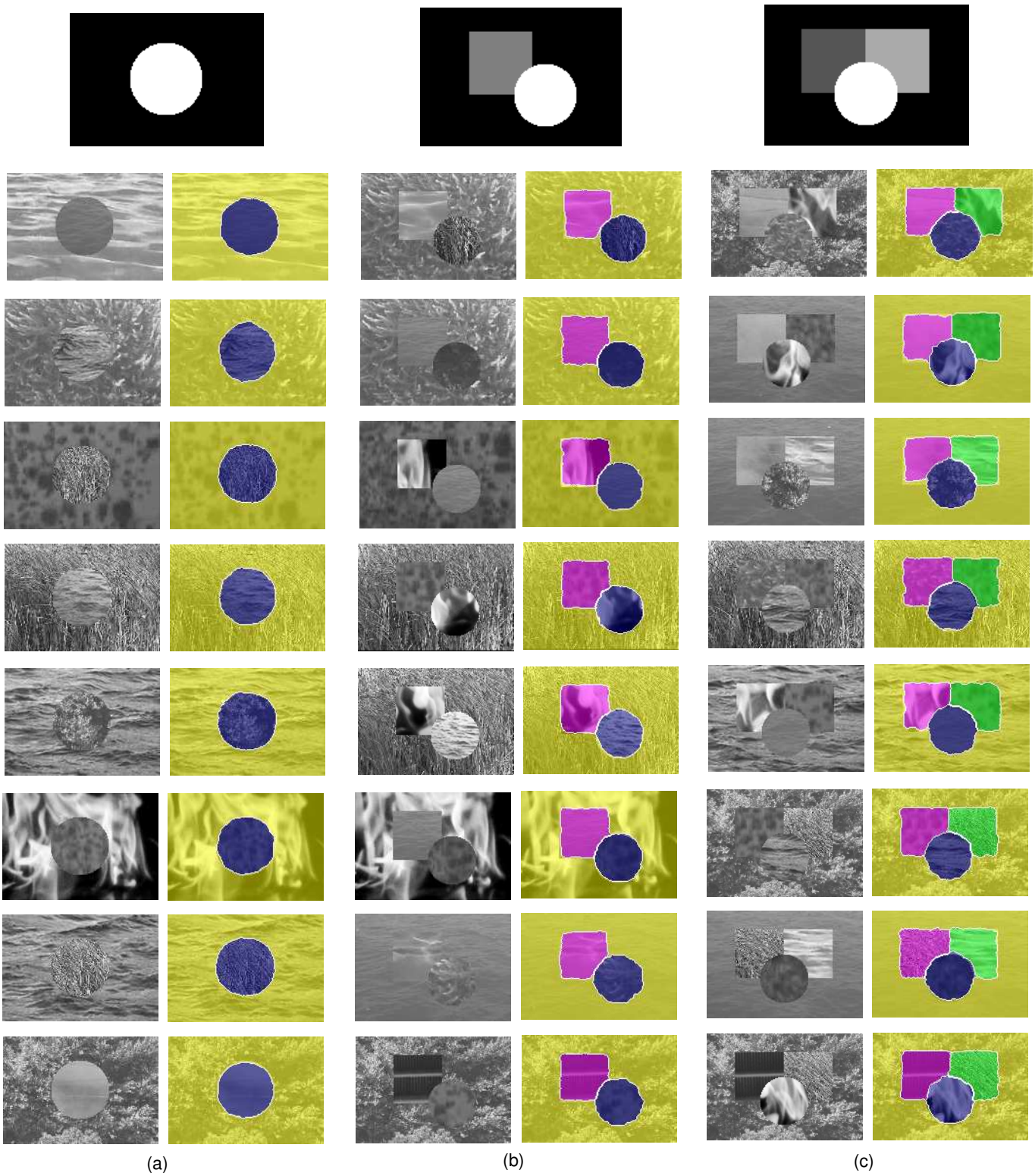

Fig. 6. Additional segmentation results obtained from the SynthDB dataset. (a) one labels, (b) two labels and (c) three labels.

[2] J. Chen, G. Zhao, M. Salo, E. Rahtu, and M. Pietikainen, "Automatic Dynamic Texture Segmentation Using Local Descriptors and Optical Flow," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 326–339, 2013.

[3] G. Doretto, D. Cremers, P. Favaro, and S. Soatto, "Dynamic texture segmentation," *in Proc. 9th IEEE International Conference on Computer Vision (ICCV)*, vol. 2, pp. 1236-1242, Oct. 2003.

[4] R. Vidal and A. Ravichandran, "Optical flow estimation and segmentation of multiple moving dynamic textures," *in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 2, pp. 516–521.

[5] A. B. Chan and N. Vasconcelos, "Modeling clustering and segmenting video with mixtures of dynamic textures, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 909–926, May 2008.

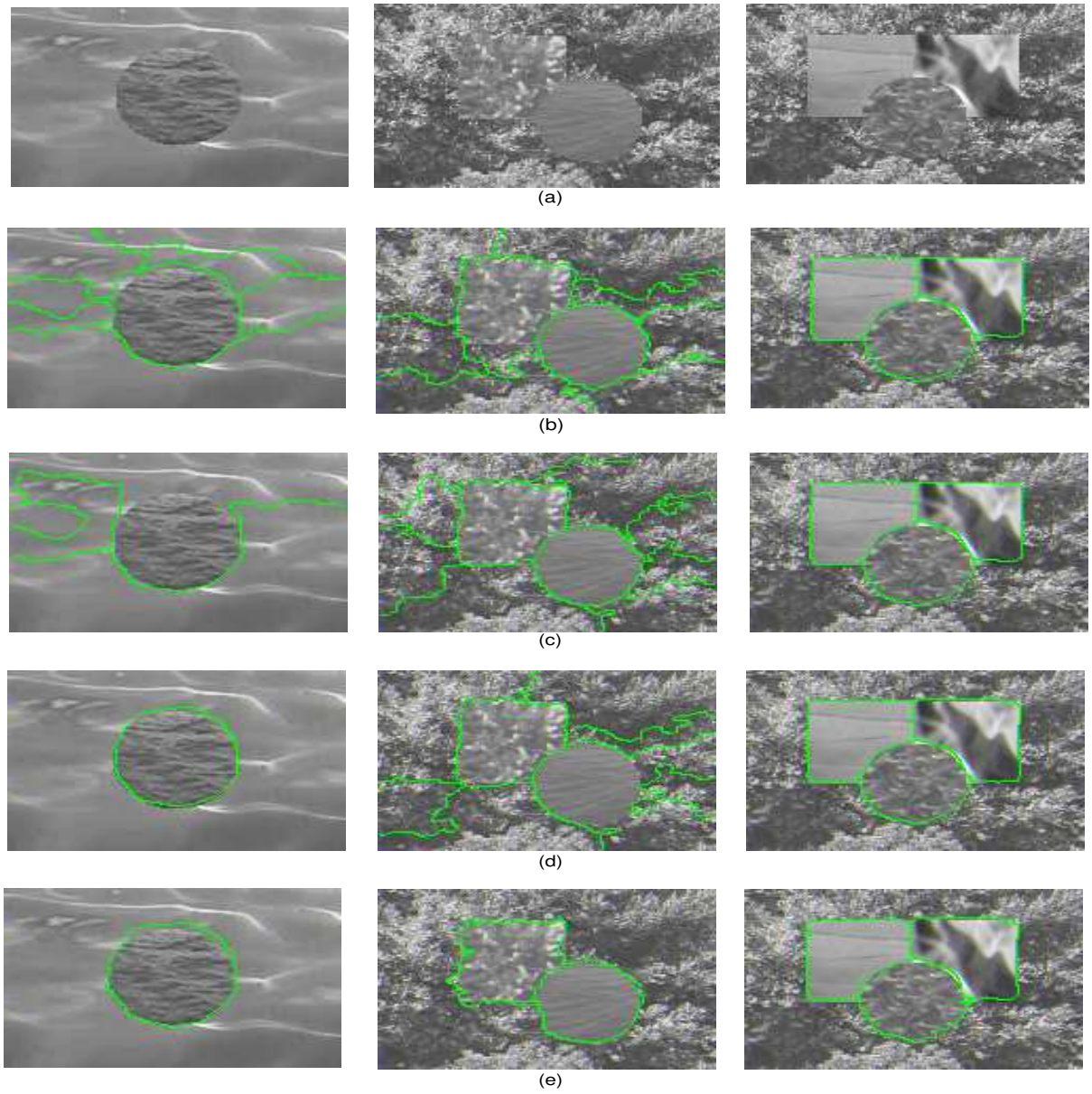

Fig. 7. Examples of segmentation results obtained by our proposed method of three videos based on different fusion criteria. (a) first frame of the video, (b) PRI, (c) VoI, (d) F-measure and (e) GCE.

TABLE III
PERFORMANCE OF THE PROPOSED METHOD USING DIFFERENT TEXTURE FEATURES ON THE SYNTHDB DATASET (PR INDEX, HIGHER IS BETTER).

ALGORITHMS	PERFORMANCE (Avg. PR)		
	99 videos 2 labels	100 videos 3 labels	100 videos 4 labels
-LAP-	0.782	0.684	0.674
-HOG-	0.771	0.692	0.695
-LPQ-	0.696	0.610	0.572
-OLBP-	**0.954**	0.823	**0.808**
-VLBP-	0.760	0.659	0.686
-ELBP-	0.953	**0.855**	0.796

[6] F. G. Fernandez, M. E. Buemi, J. M. Rodríguez and J. C. Jacobo-Berlles, "Performance of dynamic texture segmentation using GPU, " *J Real-Time Image Proc* ,vol. 11, pp. 1–9, 2016.

[7] K. Wattanachote and T. K. Shih , "Automatic Dynamic Texture Transformation Based on a New Motion Coherence Metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 10, pp. 1805–1820, 2016.

[8] G. Farnebäk, "Two-frame motion estimation based on polynomial expansion," *in Image Analysis (Lecture Notes in Computer Science)*, vol. 2749. Berlin, Germany: Springer-Verlag, 2003, pp. 363–370.

[9] T. M. Nguyen and Q. J. Wu , "An unsupervised feature selection dynamic mixture model for motion segmentation," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1210–1225, 2014.

[10] D. Teney, M. Brown, D. Kit, and P. Hall, "Learning similarity metrics for dynamic scene segmentation," *in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2084–2093.

[11] B. Cai, W. Ye and J. Zhao, "A dynamic texture based segmentation method for ultrasound images with Surfacelet, HMT and parallel computing, " *Multimed Tools Appl*, vol. 78, no. 5, pp. 5381–5401, 2019.

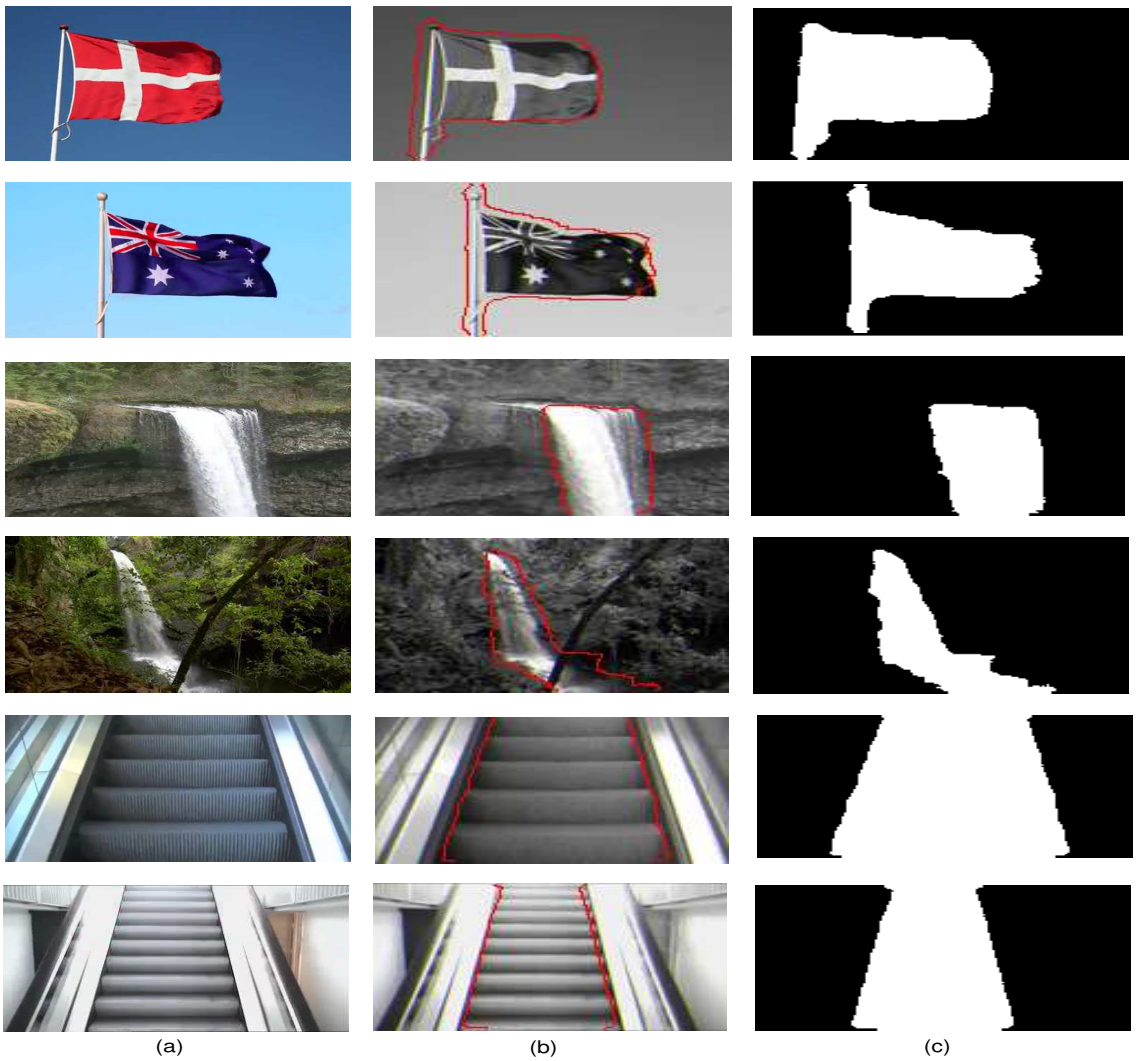

Fig. 8. Dynamic textures segmentation examples of YUP++ database (WavingFlags, Waterfall and Escalator): (a) images, (b) Segmented contours, (c) Segmented masks.

[12] S. Yousefi, M. T. M. Shalmani and A. B. Chan, "A Fully Bayesian Infinite Generative Model for Dynamic Texture Segmentation, " *ArXiv*, pp. 1–38, 2019.

[13] S. Paygude and V. Vyas, "Dynamic Texture Segmentation Approaches for Natural and Manmade Cases: Survey and Experimentation, " *Arch Computat Methods Eng*, pp. 1–13, 2018.

[14] V. Andrearczyk and P. F. Whelan,"Convolutional Neural Network on Three Orthogonal Planes for Dynamic Texture Classification," *http://arxiv.org/abs/1703.05530*, pp. 1–19, 2017.

[15] L. Khelifi and M. Mignotte, "A Consensus Framework for Segmenting Video with Dynamic Textures," *in Proc. 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2018, pp. 1–6.

[16] T. Ojala and M. Pietikäinen and T. Mäenpää, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[17] B. Lavanya and H. H. Inbarani, "A novel hybrid approach based on principal component analysis and tolerance rough similarity for face identification, " *Neural Comput and Applic*, vol. 29, no. 8, pp. 289–299, 2018.

[18] R. Marion, A. Bibal and B. Frénay, "BIR: A method for selecting the best interpretable multidimensional scaling rotation using external variables, " *Neurocomputing*, vol. 342, pp. 83–96, 2019.

[19] Q. Wang, J. Wan, F. Nie, B. Liu , C. Yan and X. Li, "Hierarchical Feature Selection for Random Projection, " *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1581–1586, 2019.

[20] B. Ella and M. Heikki, "Random Projection in Dimensionality Reduction: Applications to Image and Text Data," *in Proc. of the Seventh ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, 2001, pp. 245–250.

[21] E. Debie and K. Shafi, "Implications of the curse of dimensionality for supervised learning classifier systems: theoretical and empirical analyses, " *Pattern Analysis and Applications*, vol. 22, no. 2, pp. 519–536, May. 2019.

[22] S. P. Lloyd, "Least squares quantization in PCM, " *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–136, 1982.

[23] L. Khelifi and M. Mignotte, "A novel fusion approach based on the global consistency criterion to fusing multiple segmentations, " *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 9, pp. 2489–2502, September 2017.

[24] A. Y. Yang, J. Wright, S. Sastry, and Y. Ma, "Unsupervised segmentation of natural images via lossy data compression," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 212–225, May 2008.

[25] L. Khelifi and M. Mignotte, "GCE-based model for the fusion of multiples color image segmentations," *in Proc. 23rd IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 2574–2578.

[26] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *in Proc. 8th International Conference on Computer Vision (ICCV)*, vol. 2, July 2001, pp. 416–423.

[27] F. Destrempes, M. Mignotte, and J.-F. Angers, "A stochastic method for Bayesian estimation of hidden Markov models with application to a

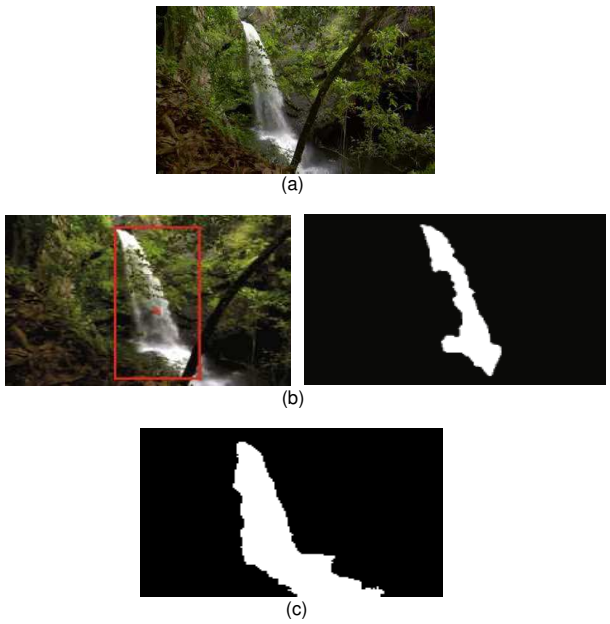

Fig. 9. An example of segmentation result obtained by our proposed method (without an initialization step) from the YUP++ dataset compared to ADAC Algorithm [32] (with an initialization step). (a) input video, (b) ADAC , (c) segmentation result based on our method.

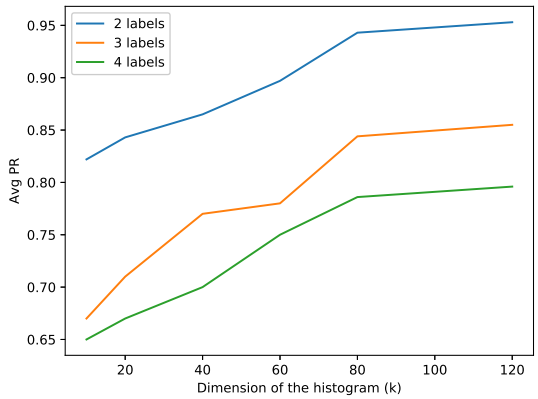

Fig. 10. Plot of the average PR obtained for each class label (of the SynthDB) as a function of the dimension of the histogram of features (k).

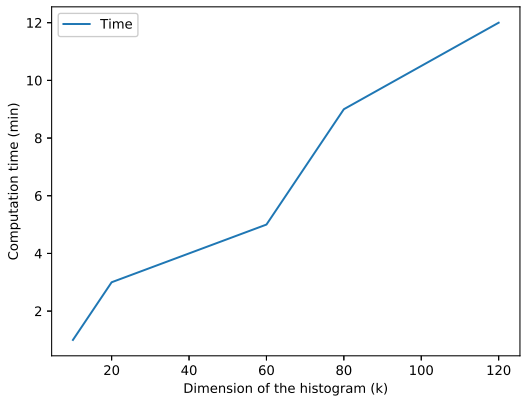

Fig. 11. Plot of the computing time as a function of the dimension of the histogram of features (k).

color model," *IEEE Transactions on Image Processing*, vol. 14, no. 8, pp. 1096–1108, 2005.

[28] L. Khelifi, I. Zidi, K. Zidi, and K. Ghedira," A hybrid approach based on multi-objective simulated annealing and tabu search to solve the dynamic dial a ride problem," *in Proc. of the International Conference on Advanced Logistics and Transport (ICALT)*, 2013, pp. 227–232.

[29] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society*, vol. 48, no. 3, pp. 259–302, 1986.

[30] L. khelifi and M. Mignotte, "Semantic image segmentation using the ICM algorithm," *in Proc. 24th IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 2017.

[31] L. Khelifi and M. Mignotte, "MC-SSM: Nonparametric Semantic Image Segmentation With the ICM Algorithm," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1946–1959, 2019.

[32] I. Bida and S. Aouat, "Dynamic Textures Segmentation and Tracking Using Optical Flow and Active Contours, " *Information Systems and Technologies to Support Learning*, editor: A. Rocha and M. Serrhini , pp. 694–704, 2019.

[33] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "A measure for objective evaluation of image segmentation algorithms, " *in Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, Jun. 2005, pp. 34–41.

[34] T. M. Nguyen and Q. J. Wu , "A Consensus Model for Motion Segmentation in Dynamic Scenes," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 12, pp. 2240–2249, 2016.

[35] A. B. Chan and N. Vasconcelos, "Layered dynamic textures, " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1862–1879, Oct. 2009.

[36] M. Mignotte, C. Collet, P. Perez, P. Bouthemy, " Hybrid genetic optimization and statistical model-based approach for the classification of shadow shapes in sonar imagery", " *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 2, pp. 129–141, Feb. 2000.