EUSIPCO 2013 1569743457

IMPROVING DYNAMIC TEXTURE RECOGNITION BY USING A COLOR SPATIO-TEMPORAL DECOMPOSITION

Rahul MOURYA and Sloven DUBOIS and Olivier ALATA and Alain TRÉMEAU

Université de Lyon, F-42023, CNRS, UMR5516, Laboratoire Hubert Curien, F-42000
Université de Saint-Étienne, Jean Monnet, F-42000, Saint-Étienne, France

ABSTRACT

The study of Dynamic Textures (DT) is a recent research topic in the field of video processing. Description and recognition of this phenomena is notoriously a difficult problem but necessary, for example, in video indexation system or video synthesis. The contribution of this paper is to show that it is possible to improve the recognition of a color DT with only a part of its information. In our approach, we propose to split a color image sequence into two components (a geometrical component and a textural component) using the Vectorial Rudin-Osher-Fatemi (VROF) model. The obtained components are used in an application of dynamic texture recognition. The experimental results clearly show that the textural part gives better recognition rates than those obtained with the geometrical part or the original video.

Index Terms— Color Dynamic Textures, Color Spatio-Temporal Decomposition, ARMA processes, recognition

1. INTRODUCTION

Textures are one of the important components of our visual world and are composed of different surfaces at different depths, orientations, with different material properties (reflectance), which are viewed under different light distributions. Some of the structures appear to be static and some appear to be in motion. The textures, which have stationarity property both spatially and temporally are referred as Temporal or Dynamic Textures (DT). The videos of processes such as waves on water surface, smoke, fire, flag fluttering in wind, a moving escalator, a walking crowd or moving vehicles observed from a certain distance, are some examples of DTs. In last three decades, the study of DT has gained popularity in both computer graphics and computer vision communities because of its vast applications such as synthesis of natural and artificial scenes in gaming and entertainment [1], video indexing/retrieval [2], video surveillance [3], background subtraction [4], tracking objects in dynamic scenes [5], *etc*.

In this paper, we address the question of DT description and recognition. This one is an active area of research in computer vision. The authors of [6] present a brief survey on description and recognition of DT. They categorize the existing approaches into five classes: methods based on optic flow [7], methods computing geometric properties in the spatio-temporal domain [8], methods based on local spatio-temporal filtering [9], methods using global spatio-temporal transforms [2], and finally, model-based methods that use estimated model parameters as features [1, 10]. Among all these approaches, we focus hereafter on the last category. More specifically, we use the Doretto's model [1] for representing image sequence because this one considers a DT as an outcome of Linear Dynamical System (LDS) that jointly involves the geometrical, photometric and dynamic features.

The main contribution that we address in this article is: is it possible to better recognize a DT with only a part of its information? To answer this question, in Section 2, we propose to decompose a color DT into geometrical and textural parts with the Vectorial Rudin-Osher-Fatimi (VROF) model [11]. In Section 3, we represent each obtained component with the linear Auto-Regressive Moving Average (ARMA) model of DT proposed by Doretto *et al.* [1]. This model considers DT as an outcome of a second-order stationary process which can be modeled as Linear Dynamical System (LDS) that evolves with time. In Section 4, we propose to use Martin distance between the Doretto's models as a tool for DT recognition. The recognition rates using just the textural component is better than using the original video or the geometrical part. Finally, Section 5 concludes this work and present future perspectives on it.

2. COLOR DYNAMIC TEXTURE DECOMPOSITION

In this section, we present the decomposition of a color DT into textural and geometrical component using Vectorial Total Variation (VTV). The main goal is to simplify the DT information for a better representation. In [11], X. Bresson and T.F. Chan propose a Vectorial Rudin-Osher-Fatemi (VROF) model for denoising color images while preserving main features such as edges. Let us consider a color image sequence $\mathbf{f}$, defined on domain $\Omega \subset \mathbb{R}^3$ with $\Omega \in [1, N_x] \times [1, N_y] \times [1, N_t]$, as follows:

$$\begin{aligned} \mathbf{f} : &\Omega \rightarrow \mathbb{R}^3 \\ &\upsilon \rightarrow \mathbf{f}(\upsilon) := (f_R(\upsilon), f_G(\upsilon), f_B(\upsilon)) \end{aligned} \tag{1}$$

with $f_R(\upsilon)$ (respectively $f_G(\upsilon)$ and $f_B(\upsilon)$) the value of voxel υ of coordinates (x, y, t) for red channel (respectively green and blue). The VROF allows to obtain two components: $\mathbf{u}$ a piecewise smooth part (geometrical) and $\mathbf{v} = \mathbf{f} - \mathbf{u}$ a residual part (textural). This model can be extended for the analysis of color video. The VROF model is given by minimizing a functional F as follows:

$$\inf_{\mathbf{u}} \left\{ F(\mathbf{u}) := \|\mathbf{u}\|_{BV(\Omega;\mathbb{R}^3)} + \frac{1}{2\lambda}\|\mathbf{f}-\mathbf{u}\|^2_{L^2(\Omega;\mathbb{R}^3)} \right\} \quad (2)$$

where $\|\mathbf{u}\|_{BV(\Omega;\mathbb{R}^3)}$ is the Vectorial Total Variation (VTV) and parameter λ controls the $L^2 - norm$ of the residual part $\mathbf{f} - \mathbf{u}$. For minimizing this functional, extended Chambolle's projection is used [11]. The extended Chambolle's projection Π on space $\lambda K_{BV(\Omega;\mathbb{R}^3)}$[1] of $\mathbf{f}$ is denoted $\Pi_{\lambda K_{BV(\Omega;\mathbb{R}^3)}}(\mathbf{f})$. It is possible to solve this projection with an iterative algorithm. For each channel $c = \{R, G, B\}$, the algorithm starts with $\mathbf{P}_c^0 = \mathbf{0}^{N_x \times N_y \times N_t}$ and for each step we have:

$$\mathbf{P}_c^{k+1} = \frac{\mathbf{P}_c^k + \delta t \nabla \left(\nabla \cdot \mathbf{P}_c^k - f_c/\lambda \right)}{1 + \delta t \sqrt{\sum_{h=\{R,G,B\}} |\nabla \left(\nabla \cdot \mathbf{P}_h^k - f_h/\lambda \right)|^2}} \quad (3)$$

until $\max\left(|\lambda \nabla \cdot \mathbf{P}_c^{k+1} - \lambda \nabla \cdot \mathbf{P}_c^k| \right) \leq r$, with $\nabla \mathbf{p}$ the spatio-temporal gradient of vector $\mathbf{p}$, $\nabla \cdot \mathbf{q}$ the spatio-temporal divergence operator of vector $\mathbf{q}$ and r is a given residue. To ensure the convergence of this iterative algorithm [11], $\delta t \leq 1/8$.

For an application of the VROF model on DT, it is necessary to extend the gradient and divergence operators to image sequence domain. For a voxel $\upsilon = (x, y, t)$ of image sequence $\mathbf{f}$, the spatio-temporal gradient is defined as follows:

$$(\nabla \mathbf{f})_{x,y,t} = \left((\nabla \mathbf{f})^x_{x,y,t}, (\nabla \mathbf{f})^y_{x,y,t}, (\nabla \mathbf{f})^t_{x,y,t} \right) \quad (4)$$

with

$$\begin{aligned}
(\nabla \mathbf{f})^x_{x,y,t} &= \begin{cases} \mathbf{f}_{x+1,y,t} - \mathbf{f}_{x,y,t} & \text{if } 1 < x < N_x \\ 0 & \text{if } x = N_x \end{cases} \\
(\nabla \mathbf{f})^y_{x,y,t} &= \begin{cases} \mathbf{f}_{x,y+1,t} - \mathbf{f}_{x,y,t} & \text{if } 1 < y < N_y \\ 0 & \text{if } y = N_y \end{cases} \\
(\nabla \mathbf{f})^t_{x,y,t} &= \begin{cases} \mathbf{f}_{x,y,t+1} - \mathbf{f}_{x,y,t} & \text{if } 1 < t < N_t \\ 0 & \text{if } t = N_t \end{cases}
\end{aligned} \quad (5)$$

In the same way, we define the spatio-temporal divergence operator of a vector $\mathbf{a} = (a^x, a^y, a^t)$ at voxel $\upsilon = (x, y, t)$ as follows:

$$\begin{aligned}
(\nabla \cdot \mathbf{a})_{x,y,t} = & \begin{cases} (a^x)_{x,y,t} - (a^x)_{x-1,y,t} & \text{if } 1 < x < N_x \\ (a^x)_{x,y,t} & \text{if } x = 1 \\ -(a^x)_{x-1,y,t} & \text{if } x = N_x \end{cases} \\
& + \begin{cases} (a^y)_{x,y,t} - (a^y)_{x,y-1,t} & \text{if } 1 < y < N_y \\ (a^y)_{x,y,t} & \text{if } y = 1 \\ -(a^y)_{x,y-1,t} & \text{if } y = N_y \end{cases} \\
& + \begin{cases} (a^t)_{x,y,t} - (a^t)_{x,y,t-1} & \text{if } 1 < t < N_t \\ (a^t)_{x,y,t} & \text{if } t = 1 \\ -(a^t)_{x,y,t-1} & \text{if } t = N_t \end{cases}
\end{aligned} \quad (6)$$

The algorithm for splitting a color DT $\mathbf{f}$ into two components, a geometrical part $\mathbf{u}$ and a textural part $\mathbf{v}$, is the following:

$$\begin{aligned}
\mathbf{u} &= \mathbf{f} - \Pi_{\lambda K_{BV(\Omega;\mathbb{R}^3)}}(\mathbf{f}) \\
\mathbf{v} &= \mathbf{f} - \mathbf{u} = \Pi_{\lambda K_{BV(\Omega;\mathbb{R}^3)}}(\mathbf{f})
\end{aligned} \quad (7)$$

The Figure 1 shows the decomposition of a color DT using the VROF model with two different values of parameter λ.

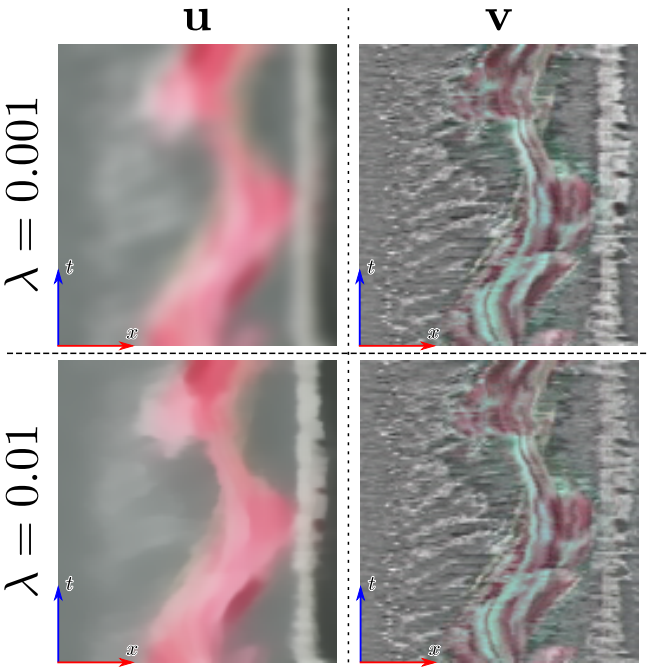

Fig. 1. Decomposition of a color DT into a geometrical part $\mathbf{u}$ and a textural part $\mathbf{v}$ with two different values of parameter λ. For a better visualization, only one temporal slice is shown. See Figure 2, for another view of the used sequence.

In the next section, we model the obtained components with the Doretto's approach.

[1] $K_{BV(\Omega;\mathbb{R}^3)}$ is the closed convex set associated to $\|.\|_{BV(\Omega;\mathbb{R}^3)}$ and defined as $K_{BV(\Omega;\mathbb{R}^3)} = \{\nabla \mathbf{p} \in L^2(\Omega;\mathbb{R}^3), \forall \mathbf{p} \in L^2(\Omega;\mathbb{R}^{3\times 3}) : |\mathbf{p}| \leq 1\}$

3. DISTANCE BETWEEN TWO DYNAMIC TEXTURES

For computing a distance between two DTs, we represent each DT with the Doretto's model and we use the obtained parameters for comparing a distance. The Doretto's approach proposed in [1] is briefly recalled here. For simplicity, we present in this section this method using $\{y_{(t)}\}_{t=1,\cdots,N_t}$ as sequence of N_t color images with $y_{(t)} \in \mathbb{R}^m$ [2]. This image sequence can be the different obtained components (**u** and **v**) or original video **f**. At each instant t, the model proposed by Doretto *et al.* represents an observed image $y_{(t)}$ of a color DT as a noisy version of ideal color image $I_{(t)} \in \mathbb{R}^m$ with an Independently Identically Distributed (IID) sequence $w_{(t)} \in \mathbb{R}^m$ (known as measurement noise) drawn from a unknown distribution.

The sequence $\{I_{(t)}\}_{t=1,\cdots,N_t}$ is a Linear Dynamic Texture (LDT) if there exist n spatial filters as the column vectors of a matrix $C \in \mathbb{R}^{m\times n}$ and hidden states $x_{(t)} \in \mathbb{R}^n$ such that $I_{(t)} = Cx_{(t)}$, and $x_{(t+1)} = Ax_{(t)} + z_{(t)}$ with $z_{(t)} \in \mathbb{R}^n$ an IID realization (known as process noise) from the unknown density, and for some choices of matrices $A \in \mathbb{R}^{n\times n}$. Thus, a LDT is a second-order stationary process that can be modeled as an ARMA process with unknown input distribution $z_{(t)}$:

$$\begin{aligned} x_{(t+1)} &= Ax_{(t)} + z_{(t)} \\ y_{(t)} &= Cx_{(t)} + w_{(t)} \end{aligned} \tag{8}$$

The matrix A is called as system dynamics or transition matrix which fully controls the transition of the states (as completely characterized by its eigen values). The matrix C is called observation matrix which describes how the hidden states are transformed into observable world.

The authors of [1] derives a closed-form procedure to learn model parameters: A, C and the noise covariance matrices. With this representation, we keep n as parameter for user: this one corresponds to the vector dimension of hidden states.

For computing the distance between two Doretto's model, different distances are available. Based on principal angle between the dynamical models, Martin in [12] proposes distance for Single Input, Single Output (SISO) linear Gaussian processes, which is also extended for multivariate case too. There are some other distances based on the subspace angles such as Finsler, Gap, and Frobenius distances, but here only Martin distance is considered because other distances do not give better results [13]. The matrices pair $\{A, C\}$ does contain most significant information for any dynamical system. To calculate the principal angles and Martin distances, only matrices pair $\{A, C\}$ is considered, neglecting the noise covariance matrices. In the next section, we firstly present our experimental protocol and secondly the obtained results.

[2] The values of the color image are reorganized in a vectorial form. m is the size of the vector and so $m = N_x.N_y.3$.

4. RESULTS

4.1. Experimental protocol

Our experimental protocol, illustrated in Figure 2, is composed of three main parts:

◦ The original videos are quite big in spatial size thus the spatial size of the video is down-sampled to 35% of the original size (720×576) by using bicubic interpolation. The down-sampled videos are decomposed into geometrical and textural components by the technique discussed in Section 2 using two different parameter values for $\lambda \in \{0.01, 0.001\}$.

◦ Doretto's models are learned separately for the down-sampled original video, and the decomposed components, by the estimation method discussed in section 3. The vector dimension of hidden states is chosen to be $n \in \{10, 20, 30, 40, 50\}$.

◦ The leave-one-out cross validation using k-nearest neighbors is used for the recognition step. Martin distances are used for comparing two models.

To evaluate our approach, we perform recognition of DT on three recently publicly available benchmarks[3]:

◦ **Alpha dataset** contains 60 image sequences divided into 3 categories *viz.* Sea, Grass, and Trees with 20 examples per categories.

◦ **Beta dataset** is more versatile, and contains 162 videos divided into 10 categories *viz.* Sea (20), Vegetation (20), Trees (20), Flags (20), Calm Water (20), Fountains (20), Smoke (16), Escalator (7), Traffic (9), and Rotation (10), where the number in the brackets represents the number of examples in each categories.

◦ **Gamma dataset** is more complex and challenging. Indeed, some classes are composed of many samples covering many cases (change in scale, orientation, *etc*). Moreover, some categories may be considered as identical, but there are two different DT phenomena (for example calm water *vs.* sea). This dataset contains 264 image sequences divided into 10 categories *viz.* Flowers (29), Sea (38), Naked trees (25), Foliage (35), Escalator (7), Calm water (30), Flags (31), Grass (23), Traffic (9) and Fountains (37).

In the next section, the recognition results of this experimental protocol are presented and discussed.

4.2. Experimental results

The overall presented results in this section were calculated in the RGB color space. With preliminary experiments, we have verified that the choice of color space (Lab, YCbCr) has a little influence on the recognition results. Indeed, we experimentally obtained almost the same results for both the synthesis and recognition applications whatever the three color spaces used.

[3] `http://projects.cwi.nl/dyntex/classification_datasets/classification_datasets.html`

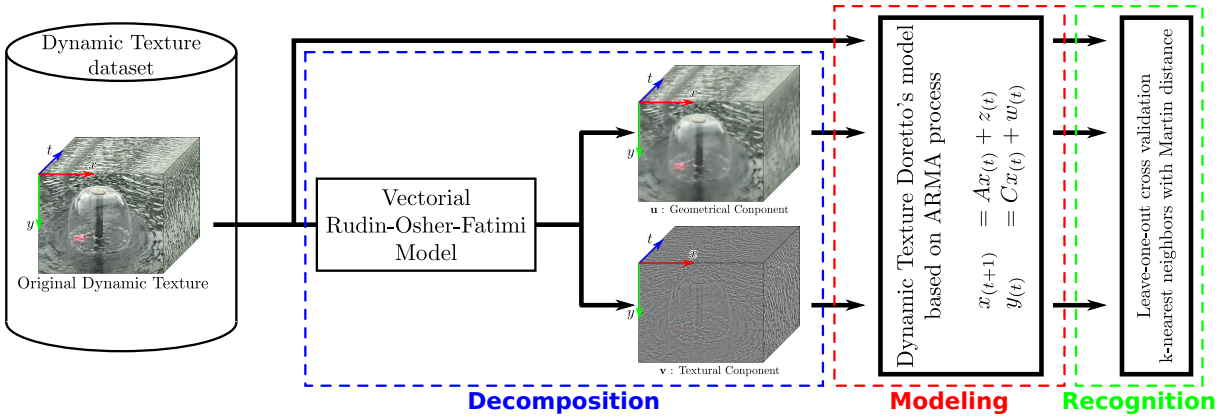

Fig. 2. Experimental protocol: (1) decomposition of a color DT, (2) Doretto's model, (3) recognition of DT.

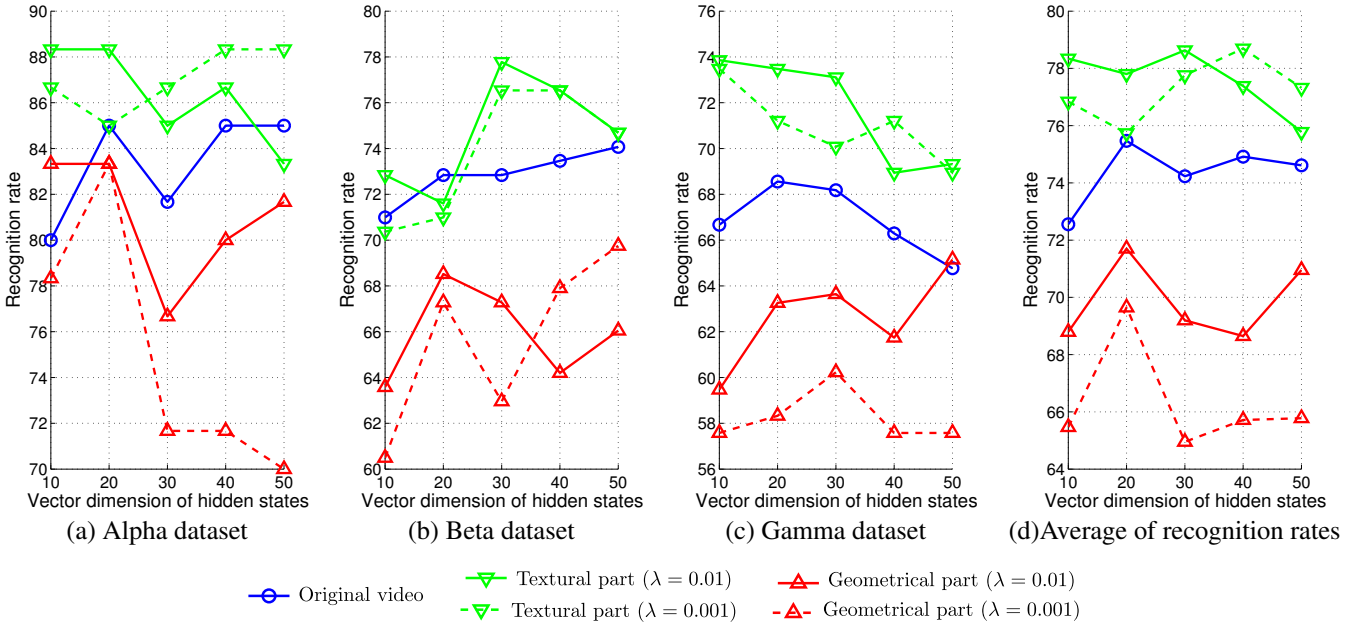

Fig. 3. (a), (b) and (c): Recognition rates for each dataset according to the vector dimension of hidden states in Doretto's approach n and to the regularization parameter λ. (d): Average of recognition rates for the three datasets for studying the parameters impact.

Figures 3.(a), 3.(b) and 3.(c) present for the three datasets, the rates of good recognition for each component (geometrical, textural or original) according to parameters λ and n (respectively the regularization parameter and the vector dimension of hidden states in Doretto's approach).

Curves (a), (b) and (c) on Figure 3 show that the textural component brings more discriminative information than the geometrical part or the original video. Indeed, in most cases, the recognition rates obtained with Doretto's model learned on textural component is better than those obtained with original video or geometrical component. These results clearly show that the sharply regions with motions (present in textural component) contain a high discriminative information. This confirms our assumption: it is of interest to decompose a complex signal for a better understanding of its information.

In most configurations, and for each dataset, our approach (with textural part) gives better recognition rates than the results in the literature. Indeed, in [2], using the coefficients of wavelet transforms, the authors obtained respectively 88%, 70% and 68% for dataset Alpha, Beta and Gamma.

In these experiments, we also study the influence of parameters λ and n. Figure 3 shows, for each vector dimension of hidden states n and for each decomposition, the average of recognition rates for the three datasets. On this curve, we can see that the previous observations are always valid.

It is also possible to better see the impact of the parameter n according to λ. Indeed, more the regularization is strong (λ decreasing), more the sharply regions and the edges are

present in textural component. For a good characterization of a rich textural part, it is necessary that the vector dimension of hidden states in Dorreto's approach is great. Curves of Figure 3.(d) illustrate clearly this point.

To conclude on the choice of parameters, it seems interesting to use a strong λ (weak regularization) with a vector dimension of hidden states sufficiently great (for example, $n = 30$). Moreover, the computation time increases drastically with n.

5. CONCLUSIONS AND PROSPECTS

In this paper, we show that it is possible to improve the recognition of a color DT with only a part of its information. Indeed, we decompose a color DT into geometrical and textural components with the VROF model. In a second part, we learn the model parameters from each part and from original video following Doretto's approach. Finally, we perform the recognition of DTs using Martin distances estimated from the parameters of the Doretto's models. The obtained recognition rates using the textural component is higher than those computed using the original video or the geometrical part. Moreover, this approach has an advantage that one does not need to explicitly select the key regions in a scene for recognition because the textural component dominantly contains only the textural regions in the video.

This work allows many prospects. Indeed, we have used here the VROF model for decomposing the video. It may be interesting to decompose the image sequence in a different way. For example, with a decomposition model that respects more edges [14] or with a decomposition that can split the textural component between different parts (spatial, temporal and spatio-temporal texture) we could obtain more discriminative parts. We also think about characterizing the components in different ways. For example, it is possible to use an other descriptor than Doretto's model parameters. Finally, for classification step, other algorithms as Support Vector Machines (SVM) could be used to achieve better classification rates.

6. REFERENCES

[1] G. Doretto, A. Chiuso, Y.N. Wu, and S. Soatto, "Dynamic Textures," *Computer Vision*, vol. 51, pp. 91–109, 2003.

[2] S. Dubois, R. Péteri, and M. Ménard, "Characterization and Recognition of Dynamic Textures based on 2D+T Curvelet Transform," *in Press Signal, Image and Video Processing*, 2013.

[3] W. Phillips, M. Shah, and N.V. Lobo, "Flame Recognition in Video," *Pattern Recognition Letters*, vol. 23, pp. 319–327, 2002.

[4] A.B. Chan, V. Mahadevan, and N. Vasconcelos, "Generalized Stauffer-Grimson Background Subtraction for Dynamic Scenes," *Machine Vision and Application*, vol. 22, pp. 751–766, 2011.

[5] R. Péteri, "Tracking Dynamic Textures using a Particle Filter Driven by Intrinsic Motion Information," *Machine Vision and Applications*, vol. 22, pp. 781–789, 2011.

[6] D. Chetverikov and R. Péteri, "A Brief Survey of Dynamic Texture Description and Recognition," in *CORES 2005*, pp. 17–26.

[7] R.C. Nelson and R. Polana, "Qualitative Recognition of Motion using Temporal Texture," *Computer Vision and Image Understanding*, vol. 56, pp. 78–89, 1992.

[8] K. Otsuka, T. Horikoshi, S. Suzuki, and M. Fujii, "Feature Extraction of Temporal Texture Based on Spatiotemporal Motion Trajectory," in *ICPR 1998*, pp. 1047–1051.

[9] G. Zhao and M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *Pattern Analysis and Machine Intelligence*, vol. 29, pp. 915–928, 2007.

[10] M. Szummer and R.W. Picard, "Temporal Texture Modeling," in *ICIP 1996*, pp. 823–826.

[11] X. Bresson and T. F. Chan, "Fast Dual Minimization of the Vectorial Total Variation Norm and Applications to Color Image Processing," *Inverse Problems and Imaging*, vol. 2, no. 4, pp. 455–484, 2008.

[12] R.J. Martin, "A metric for ARMA processes," *Signal Processing*, vol. 48, no. 4, pp. 1164–1170, 2000.

[13] P. Saisan, G. Doretto, Y.N. Wu, and S. Soatto, "Dynamic Texture Recognition," in *CVPR 2001*, pp. 58–63.

[14] A. El Hamidi, M. Ménard, M. Lugiez, and C. Ghannam, "Weighted and extended total variation for image restoration and decomposition," *Pattern Recognition*, vol. 43, no. 4, pp. 1564 – 1576, 2010.