FASTensor: A tensor framework for spatiotemporal description

Virgínia F. Mota
Colégio Técnico
Universidade Federal de Minas Gerais
Email: virginiaferm@dcc.ufmg.br

Jefersson A. dos Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Email: jefersson@dcc.ufmg.br

Arnaldo de A. Araújo
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Email: arnaldo@dcc.ufmg.br

***Abstract*—Spatiotemporal description is a research field with applications in various areas such as video indexing, surveillance, human-computer interfaces, among others. Big Data problems in large databases are now being treated with Deep Learning tools, however we still have room for improvement in spatiotemporal handcraft description. Moreover, we still have problems that involve small data in which data augmentation and other techniques are not valid. The main contribution of this Ph.D. Thesis[1] is the development of a framework for spatiotemporal representation using orientation tensors enabling dimension reduction and invariance. This is a multipurpose framework called Features As Spatiotemporal Tensors (FASTensor). We evaluate this framework in three different applications: Human Action recognition, Video Pornography classification and Cancer Cell classification. The latter one is also a contribution of this work, since we introduce a new dataset called Melanoma Cancer Cell dataset (MCC). It is a small data that cannot be artificially augmented due the difficulty of extraction and the nature of motion. The results were competitive, while also being fast and simple to implement. Finally, our results in the MCC dataset can be used in other cancer cell treatment analysis.**

I. Introduction

Spatiotemporal data usually contains the states of an object, an event or a position in space over a period of time. This data can be created from videos or multitemporal images (sequences of images that are combined depending on the purpose). An event in a spatiotemporal dataset describes a spatial and temporal phenomenon that may happen at a certain time and location.

In order to learn useful information regarding these events, computational systems generally use combinations of different features representing visual elements from the scene, such as color, texture, salient points, apparent motion, trajectories, *etc*. Those visual patterns provide information on the two-dimensional and/or three-dimensional structure of the scene, shape and trajectory of objects and the activity that is going on. Therefore, this visual information of still and moving images is the key for tasks such as: video compression [1], object tracking [2], video segmentation [3], video surveillance [4], video and multitemporal image classification [5], cell shape classification [6].

All those tasks that work with moving pictures need to be represented using not only spatial characteristics, but spatiotemporal description. It is a challenging application as we have continuous and discrete changes on the scene being influenced locally and globally, both in time and space. Let us take as an example two actions from the video dataset KTH [7]: Jogging and Walking. For both, we have a person doing the action in a homogeneous background that involves moving their feet. However, those actions are slightly different according to their velocities. So, we have to take into account the shape of the movement, the coherence through time, the velocity between frames. This exemplifies the challenges of extracting semantic information from elements in a scene that do not intrinsically possess semantic meaning, but instead are encoded as sequential numerical matrices with temporal variations.

In this thesis, we address the spatiotemporal feature representation problem applied to video and multitemporal image classification. Many works tackle this problem following three steps: handcrafted feature extraction, descriptor creation, and classification. We are mainly interested in the first two steps: *feature extraction* and *descriptor creation*.

We classify the existing methods based on the type of features: Shallow and Deep Learning (DL) methods. Shallow methods are categorized into the following classes: *1*. Low-level approaches with handcrafted features; *2*. Bag-of-feature (BOF) representations or middle-level approaches. Deep Learning-based approaches share similar procedures: patch sampling, feature description/learning and classification [8]. Nowadays, image and video classification problems in large databases are being treated with Deep Learning tools [9], [10].

However, deep architecture models suffer from over-fitting problems when there is a small amount of training data. There are methods to overcome this problem, such as data augmentation, transfer learning, data generation, among others. But coherent approaches for moving pictures are still in their infancy as well as adding the temporal information on a deep architecture [9]. Hence, this is still an open problem in literature.

For the shallow methods, the handcrafted feature extraction starts by a preliminary dimension reduction since some point based motion indicator, usually intensity gradient, is coded in a compact form. Feature examples include Histogram of Gradients (HOG),Histogram of Optical Flow (HOF), basis projections, and other Optical Flow (OF) based features. In

[1]This work relates to a Ph.D. Thesis

most works of the literature, these features are associated with Scale-Invariant Feature Transform (SIFT), Speeded Up Robust Feature (SURF) or Spatio Temporal Interest Points (STIP) descriptors [11].

The description creation step uses the extracted features to provide the video signature, using a single type or a combination of features. The most used method for shallow methods is the canonical BOF. We discuss other methods to create the video signature. Using the idea of coding features into orientation tensors, we are able to aggregate them in order to represent the temporal evolution.

Different from the shallow methods, Deep Learning-based approaches do not usually work with handcrafted feature extraction. A deep feature is the consistent response of a unit within a hierarchical model to an input, where this response contributes to the model decision. A feature could be considered deeper than another depending on where the unit is positioned alongside the hierarchical structure of the model [9].

In this thesis, we work with handcrafted and deep feature extraction, thus the classification method follows a shallow approach. The video classification step is used to evaluate the descriptors created. We work with three spatiotemporal representation tasks: Human Action Recognition, Video Pornography classification and Cancer Cell classification.

The main contribution of this work is the development of a novel spatiotemporal description framework (Features As Spatiotemporal Tensors – *FASTensor*) using orientation tensors, enabling dimensionality reduction and invariance according to the feature. In order to evaluate the framework in other distinct and challenging scenarios than the traditional computer vision tasks, we also present a new open labeled dataset for melanoma cancer cell classification. It is a small dataset, called Melanoma Cancer Cell dataset (MCC)[2], that cannot be artificially augmented due the inherent difficulties in the acquisition process and its particular nature. Furthermore, there are no similar open datasets in the literature, to the best of our knowledge. This opens the discussion on how we can we learn with small datasets. Our proposed method and experiments show that the method can be used in other cancer cell treatment analysis.

II. FASTensor: Features As Spatiotemporal Tensors

An Orientation Tensor Framework for spatiotemporal description can be modeled as shown in Figure 1.

This framework can be used in videos or multi-temporal images with temporal dimension n. The orientation tensor T_v created from each feature vector $\hat{\mathbf{v}}$ with mean μ will be accumulated for each image/frame i in order to represent the covariance of it, as in:

$$T = \sum_{i=1}^{n} T_v = \sum_{i=1}^{n} (\hat{\mathbf{v}}_i - \mu)(\hat{\mathbf{v}}_i - \mu)^T. \quad (1)$$

[2]https://tiny.cc/mcc-dataset

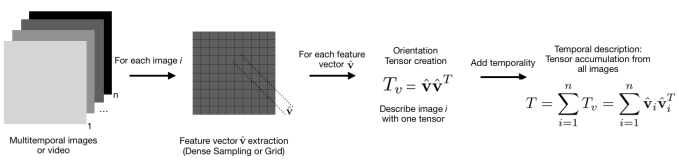

Fig. 1. An orientation tensor framework for temporal description created from the feature vector $\hat{\mathbf{v}}$ extracted with grid or dense sampling.

The features can be extracted with dense sampling or grid. Then, the accumulation through time will provide the temporal description for the video or for multi-temporal images. It is important to note that in each step, a normalization of the orientation tensor may be needed as the number of feature vectors from each image or frame could vary along time.

Figure 2 shows a two-dimensional example for the framework to better explain how the orientation tensor carries more information than the feature vector. Visually, instead of just having a vector representing the trend, we have the ellipsoid, carrying all the uncertainties and covariance measures of the features. Figure 2 shows an example of a movement tendency using a HOG feature of a person walking on a homogeneous background. With the orientation tensor, we can capture not only what happens in this scene, but how we begin to deform the ellipsoid so that it carries the whole tendency of the HOG. The geometric representation is made in three dimensions to facilitate the understanding of the problem. We will show in the following applications how this change can significantly improve video classification.

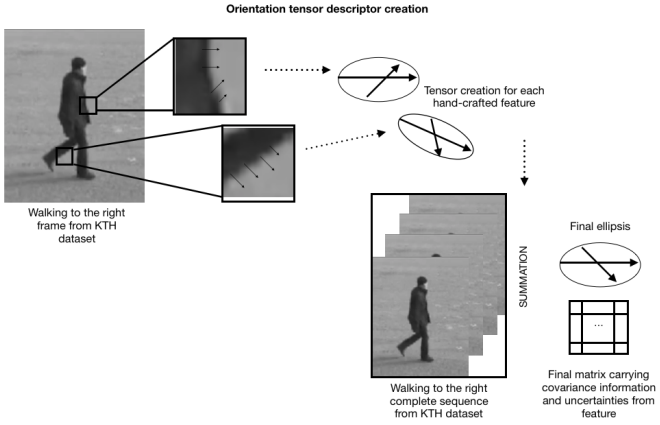

Fig. 2. Geometric example of the FASTensor framework for temporal description created from a feature vector $\hat{\mathbf{v}}$ (handcrafted) extracted with grid sampling. The final descriptor is a matrix $n \times n$, where n is the dimension of the feature, that carries the covariance and uncertainties from the features. The image is an example of walking action from KTH dataset [7].

Therefore, orientation tensors can be used as compact spatiotemporal representations, enabling dimension reduction and invariance according to the feature used to create them. They will capture the covariance information from the feature vector adding more information to the descriptor.

Given the mathematical framework and the *FASTensor* depicted in Figure 2, we can describe our proposed method with the following pseudo-algorithm:

```
(1) Input: Video or Multi-temporal images
(2) for each image i in Input:
(3)     // m features with dimension n
(4)     v[m] = feature_extraction();
(5)     for each feature vector j in v[m]:
(6)             T_j = matrix_multiplication(v[j]);
(7)             T_i = T_i + T_j;
(8)     normalize(T_i);
(9)     T_input = T_input + T_i;
(10) Output: T_input, a matrix of n x n
```

Fig. 3. Pseudo-algorithm for the FASTensor framework.

The core of *FASTensor* is the computation of orientation tensor for each feature vector $\hat{\mathbf{v}}$. This is achieved with a matrix multiplication (line 6 of Figure 3), therefore, we have a complexity of $O(n^3)$, where n is the dimension of the feature vector $\hat{\mathbf{v}}$. In one frame we can have m feature vectors depending on the type of extraction. In the worst case, m is the number of pixels of the frame (dense sampling in lines 4 and 5 of Figure 3). Finally, the input has f frames (line 2 of Figure 3). Again, in the worst case, we use all frames from input.

The final complexity in the worst case for the *FASTensor* is $O(f \times m \times n^3)$, where f is the number of frames, m is the number of feature vector per frame and n is the dimension of the feature vector. In terms of time, as our method is feature dependent, we need to add the complexity of the feature extraction in $O(f \times m \times n^3)$.

Therefore, we have a complexity cubic growth in relation to the size of the feature, but a linear growth in relation to the number of features per frame and number of frames. So, we can reduce the computation time just by using less frames and other feature sampling instead of dense sampling.

The limitations of this framework are that it carries only global information from each image and it is very dependable of the used feature. Thus, for describing several information from the same image sequences the method may not be eligible. For example, the ellipse depicted in Figure 2 can become a circle (or sphere in three dimensions), not carrying any main tendency information. That is, the tensor becomes isotropic.

III. Experiments

We work with three spatiotemporal representation tasks: Human Action Recognition (Section III-A), Video Pornography classification (Section III-B) and Cancer Cell classification (Section III-C).

Setup: The experiments used three handcrafted features: HOG; HOF; and, the concatenation of both of these features (HOGHOF).

Transfer Learning using pretrained DNNs has also become a common practice for Computer Vision applications with the dawn of very large labeled datasets such as ImageNet [12] and Pascal VOC [13]. Therefore we also tested the performance of tensors on the task of adding temporal consistency on feature vectors generated by activations in pretrained Convolutional Neural Networks (CNNs) [14], as these models are originally suited only for static images. Activations at the end of four distinct residual blocks in a pretrained ResNet were used as both raw features for classification and as inputs for FASTensors and compared at Section III-B and will be henceforth named as follows: ResNet-50 (1); ResNet-50 (2); ResNet-50 (3); and, ResNet-50 (4).

The pretrained DNNs were acquired and implemented using the PyTorch framework and the torchvision pretrained model for the ResNet-50 [15].

We used Support Vector Machines (SVMs) as inference models for the classification tasks and compared the accuracy metric of baselines. Feature extraction modules in this work were implemented using the skimage framework, while SVM and validation procedure were coded using the sklearn library. The core of the FASTensor approach uses the NumPy and SciPy libraries. We present the experimental protocol in the following sections for pornography and cancer cell datasets.

A. Human Action Recognition

Our experiments used three benchmark datasets: KTH [7], UCF11 [16] and Hollywood2 [17]. Table I shows the comparison of our works in those datasets. Table II summarizes the results for KTH dataset for global appearance and motion based descriptors.

TABLE I
A comparison of all works in three benchmark datasets: KTH, UCF11 and Hollywood2. Recognition rate in percentage for each of our works.

KTH		**UCF11**		**Hollywood2**	
[18]	93.3	[19]	75.4	[19]	40.3
[20]	93.2	[20]	72.7	[20]	40.3
[19]	92.5	[21]	68.9	[21]	34.0
[21]	92.0	[22]	57.8	[22]	15.0
[22]	87.8				
[23]	86.6				

TABLE II
Recognition rates in percentage for KTH dataset using bag-of-feature based methods and our approaches. *Indicates leave-one-out protocol.

Local descriptors		Trajectories		Relationship Modeling	
[24]	95.6	[25]	97.4	[26]	98.2
[27]	94.8*	[28]	95.3	[29]	94.5
[30]	93.9	[31]	94.2	[32]	94.5
[33]	93.8				
		Tensor			
		[34]	94.2		
		Our approaches			
		[18]	93.3		
		[20]	93.2		
		[19]	92.5		
		[21]	92.0		
		[22]	87.8		
		[23]	86.6		

For UCF11, the best results are by [28] with 85.4% and by [30] achieving 75.8%. We see that for more challenging

datasets, the best results are still with [26] and [28]. Note that our best result in UCF11 is 75.4% for [19] which models the temporal evolution of HOG with orientation tensors. Thus, using only one type of feature, we achieved a recognition rate very close to a bag-of-feature technique.

A similar result is achieved for Hollywood2 dataset. Hollywood2 dataset is the most challenging, and has been collected from Hollywood movies. Table III summarizes the recognition rates for Hollywood2 dataset.

TABLE III
RECOGNITION RATES IN PERCENTAGE FOR HOLLYWOOD2 DATASET USING BAG-OF-FEATURES BASED METHODS AND OUR APPROACHES.

Local descriptors		Trajectories		Relationship Modeling	
[30]	53.3	[35]	62.5	[29]	50.9
[24]	47.7	[28]	59.9		
		Tensor			
		[36]	59.5		
		[34]	57.6		
		Our approaches			
		[19]	40.3		
		[20]	40.3		
		[21]	34.0		
		[22]	15.0		

All those results were achieved with other shallow approaches. When compared to Deep Learning-based techniques, those three datasets are already deprecated. Table IV shows the best results for KTH, UCF11 and Hollywood2 using state-of-the-art deep learning-based approaches.

TABLE IV
BEST RESULTS FOR KTH, UCF11 AND HOLLYWOOD2 USING STATE-OF-THE-ART DEEP LEARNING-BASED APPROACHES.

Dataset	Recognition Rate
KTH [37]	98.67%
UCF11 [38]	93.77%
Hollywood2 [39]	78.50%

Literature in Human Action Recognition has moved to more difficult datasets as HMDB51 [40] and Sports-1million [9]. Those datasets have more heterogeneous actions, more videos and even semantic context as smiling and laugh. Thus, for human action recognition we found a barrier and we are not able to compete with deep learning methods.

B. Video Pornography Classification

The Pornography-800 Dataset created by Avila et al. [41], contains nearly 80h of 400 pornographic and 400 non-pornographic videos. Concerning the pornographic material, the dataset is very assorted, including both professional and amateur content[3]. Moreover, it depicts several genres of pornography, from cartoon to live action, with diverse behavior and ethnicity. With respect to non-pornographic content, they are general-purpose video networks, with difficult cases like sumo, swimming, beach scenarios (i.e., words associated to skin exposure).

[3]https://sites.google.com/site/pornographydatabase/

The baseline results for this dataset are presented in Table V extracted from [41]. They preprocessed the dataset by segmenting videos into shots. On average there are 20 shots per video. A key frame (middle frame) is selected to summarize the content of the shot into a static image. As low-level local descriptor, they employed HueSIFT [42], a SIFT variant including color information. The 165-dimensional HueSIFT descriptors are extracted densely every six pixels. The same vocabulary M constructed by k-means clustering algorithm, with M fixed as 256, is used for the standard BoF and the BossaNova method [41].

For classification, they used a 5-fold cross-validation to tune the best C parameter for a SVM classifier. They reported the image classification performance by using the mean Average Precision (mAP), and the video classification by accuracy rate, where the final video label is obtained by majority voting over the images. It is interesting to note that for both reported methods, the video classification scores are inferior to the image classification scores. That can be explained by the fact that some pornographic videos have the additional difficulty of having very few shots with pornographic content (typically one or two takes among several dialog shots or cut scenes).

TABLE V
BASELINE FOR THE PORNOGRAPHY-800 DATASET USING STANDARD BAG-OF-FEATURES AND BOSSANOVA. COMPARED RESULTS FROM HANDCRAFTED FEATURES AND THE FASTENSOR FOR THE PORNOGRAPHY-800 DATASET. WE USED A DENSE SAMPLING EXTRACTION WITH FIXED NUMBER OF BINS, HOG WITH SIXTEEN BINS (EIGHT FOR EACH FRAME IN A PAIR), HOF WITH EIGHT BINS, AND HOGHOF WITH TWENTY-FOUR BINS. RESULTS FOR THE FASTENSORS FOLLOWED BY † REPRESENT ACCURACIES THAT WERE SIGNIFICANTLY IMPROVED BY THE PROPOSED APPROACHES IN COMPARISON WITH USING RAW FEATURES.

Method		Accuracy (%)
Baselines	**BoF**	83 ± 3
	BossaNova [41]	89.5 ± 1
	Caetano et al. [43]	92.4 ± 1
	TRoF [44]	95 ± **
	ACORDE-50* [10]	94.8 ± 2
	ACORDE-101* [10]	95.6 ± 1
Raw Features	**HOG**	82.16 ± 0.54
	HOF	77.20 ± 0.31
	HOGHOF	88.12 ± 0.56
	ResNet-50 (1)	91.34 ± 0.28
	ResNet-50 (2)	92.25 ± 0.14
	ResNet-50 (3)	94.73 ± 0.73
	ResNet-50 (4)	94.75 ± 0.38
FASTensors	**HOG**	85.32 ± 0.31 †
	HOF	84.18 ± 0.18 †
	HOGHOF	93.28 ± 0.36 †
	ResNet-50 (1)	93.50 ± 0.12 †
	ResNet-50 (2)	93.49 ± 0.14 †
	ResNet-50 (3)	**96.45 ± 0.24** †
	ResNet-50 (4)	96.25 ± 0.25 †

We compared our results with the accuracy from the baseline. We used the same division protocol from the baseline as SVM protocol. We compared three handcrafted features vastly used in video description: HOG, HOF and the combination of both HOGHOF. We used a dense sampling extraction with the fixed number of bins, HOG with sixteen bins, HOF with eight bins, and the HOGHOF with twenty-four bins. The

results comparing the baseline, handcrafted features and the FASTensor are depicted in Table V.

C. Cancer Cell Classification

One of the contributions of this work is a new open multitemporal image dataset: The Melanoma Cancer Cell dataset (MCC)[4]. This dataset was created in colaboration with the Biology Insitute of *Universidade Federal de Minas Gerais*. It provides better understanding of the cancer cell migration and anti-migration promoted by specific drugs [45], classifying in treated and untreated cell, being possible to characterize phenotypic and morphologic drug effects [46]. Therefore, allowing to elucidate some intrinsic biological mechanisms of cancer cell, particularly understanding the tissue invasion and metastases formation.

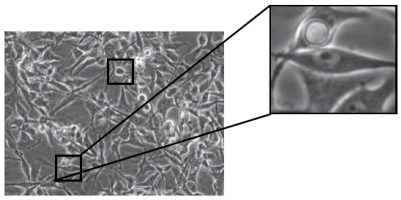

Fig. 4. Example of cells from the melanoma cancer cell dataset. Two example cells are marked with a black bold square around its nucleoid. On the right we have a zoom on one of them.

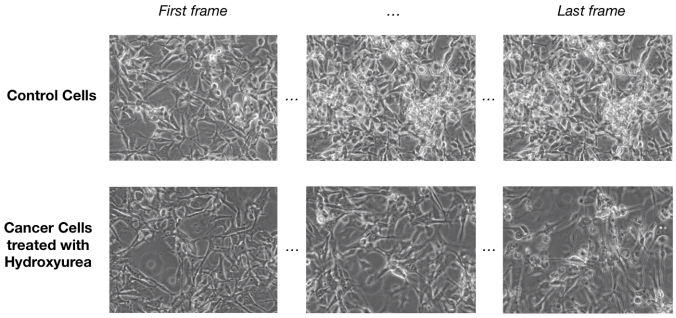

Fig. 5. The melanoma cancer cell dataset composed by 69 image sequences of control melanoma cells and 69 image sequences for cells treated with hydroxyurea. On the left, we see the evolution of melanoma cancer cells through time. On the right, we see the cells treated with hydroxyurea. It is easy to see how the number of cells increases without any treatment.

This dataset has two conditions of long-term culture of metastatic murine melanoma B16F10 cells in Roswell Park Memorial Institute (RPMI) medium (supplemented with 10% Fetal Bovine Serum, Streptomycin 10 mg/mL and Penicillin 10,000 Units/mL). First of all, B16F10 was plated (5x104 cells/mL) in a 35mm polystyrene dish and, after 24h, exposed to hydroxyurea (30mM) or only medium (control group). Then, cells were placed in BioStation IM-Q inverted microscope (Nikon) and images from 69 fields were acquired over 24 hours by a high sensitivity cooled charge-coupled device (CCD) camera (40x objective). At the end, the final database resulted in 69 image sequences with 95 frames with a spatial resolution of 640x480 pixels and duration of one minute. Figure 4 presents a frame example with two marked cells to show what is the subject of this dataset. For this dataset, image sequences are multitemporal images.

Hydroxyurea is a non-alkylating antineoplastic that selectively inhibits ribonucleoside diphosphate reductase, an enzyme required to convert ribonucleoside diphosphates into deoxyribonucleoside diphosphates, thereby preventing cells from leaving the G1/S phase of the cell cycle. In B16F10 cells, inhibition of migration by hydroxyurea starts from 1uM reaching maximum effect at 30uM without increasing cell death [45].

In the control cell image sequences we see that cells increase the number and the velocity. When hydroxyurea is applied, the number of cells and the velocity decrease over time. Thus, it is interesting to analyze how a spatiotemporal descriptor can be used to discriminate the treated cells from the control cells in order to automate the process and help us better understand the phenomena. Figure 5 shows an example of a sequence from the dataset. On the left we see the evolution of melanoma cancer cells through time. On the right we see the cells treated with hydroxyurea. With the last frame, it is easy to see how the number of cells increases without any treatment.

To evaluate the results of our experiments, we applied a 5x2-fold protocol. It consists of randomly splitting the MCC video dataset five times into two folds, balanced by class. In each time, training and testing sets were switched and consequently five analysis for every model employed were conducted.

The baseline was computed with a dense extraction of the three handcrafted features HOG, HOF and HOGHOF. The results are depicted in Table VI. It can be observed that our assumption that a video descriptor could discriminate the control cells from the cancer cells is a fact, for all handcrafted features we achieved an accuracy greater than 80%.

TABLE VI
BASELINE HANDCRAFTED FEATURES FOR THE MELANOMA CANCER CELL DATASET. WE USE A DENSE SAMPLING EXTRACTION WITH THE FIXED NUMBER OF BINS, HOG WITH SIXTEEN BINS, HOF WITH EIGHT BINS, AND THE HOGHOF WITH TWENTY-FOUR BINS.

Method	Accuracy (%)
HOG 16 bins	81.22 ± 0.14
HOF 8 bins	92.2 ± 0.62
HOGHOF 24 bins	96.9 ± 0.24

TABLE VII
FASTENSOR RESULTS FOR THE MELANOMA CANCER CELL VIDEO DATASET. ALL RESULTS ARE STATISTICALLY SIGNIFICANT BETTER THAN THE BASELINE.

Method	Accuracy (%)
HOG 16 bins	89.58 ± 0.30
HOF 8 bins	95.69 ± 0.15
HOGHOF 24 bins	99.78 ± 0.34

IV. CONCLUSION

In this thesis, we proposed an orientation tensor framework for video description called Features As Spatiotemporal Tensors (*FASTensor*). The orientation tensor created from each

[4]https://tiny.cc/mcc-dataset

feature vector is accumulated for each image/frame. The accumulation through time provides the temporal description for the video or for multi-temporal images. We showed the mathematical fundamentals and the proof of context for the framework.

We evaluated the FASTensor in three different video classification tasks: Human Action Recognition, Video Pornography classification and Melanoma Cancer Cell classification, to which we contribute with a new dataset.

Our experiments confirmed that the incorporation of covariance information from the features led to more effective video classification in different applications. This was shown with raw features HOG, HOF and HOGHOF, and deep features pretrained on a ResNet-50. In comparison with the state-of-the-art, our framework yielded better results.

For the Human Action Recognition task, it was possible to create a simple descriptor using orientation tensors that could maintain balance between size, computer complexity and recognition rate. However, the big limitation of our method is the number of actions that can be performed in one scene. Thus, for more complex video datasets we were not able to achieve competitive results, as the orientation tensor has a bigger tendency to become isotropic, that is, not have main direction information.

For the Video Pornography classification task, the FASTensor achieved the best results for the Pornography-800 and a competitive result for the Pornography-2k. In fact, this application is more suitable to work with orientation tensor, as the probability to become isotropic is inferior.

The Melanoma Cancer Cell (MCC) dataset provides better understanding of the cancer cell migration and anti-migration promoted by specific drugs, classifying in treated and untreated cell, being possible to characterize phenotypic and morphologic drug effects. This dataset showed that FASTensor can be used in very different applications. Moreover, the framework can be used in other cancer cells treatment analysis.

With our results we can, therefore, confidently assert that FASTensor comprise the new state-of-the-art for video classification in the Pornography-800 dataset and for the Melanoma Cancer Cells dataset. For Human Action Recognition, we could also achieve competitive results. Therefore, orientation tensors carry more discriminative information than the feature vector itself, showing how robust is our method.

This thesis established the theoretical fundamentals for the orientation tensor framework, furnished a statistical analysis and was able to test the FASTensor in different applications.

As future work, we want to test other drugs in cancer cells and automate the analysis. We will investigate what more can be extracted with orientation tensors for this application, like motion tendency, cell density, among others. We also want to analyze other applications that are suitable for FASTensor in medical imaging, remote sensing, surveillance, among other spatiotemporal tasks.

Furthermore, we will analyze the FASTensor as a descriptor creator not only for handcrafted features and deep learning features. We already saw the improvement for Pornography classification. We believe that we can improve the results adding temporal information without the overhead of a very deep architecture for video classification with more studies in other spatiotemporal applications. One idea is to add a layer in a CNN approach that creates tensors to add temporal information to the neural network.

A. Publications

This research produced the following published papers as contribution to the literature in spatiotemporal representation:

- Journals: [47] (Under Review), [20]
- Book Chapters: [48]
- Conferences: [49], [19]

This thesis also contributed to:

- Journals: [50], [23]
- Conferences: [51], [52], [53], [54] (Best Paper of Workshop on Vision-based Human Activity Recognition), [55], [18]
- Summer School Participation: ENS/INRIA Visual Recognition and Machine Learning Summer School. Paris,France, 22-26 July 2013. Poster presentation based on [21].

ACKNOWLEDGMENT

Authors would like to thank to UFMG, CAPES, CNPq for funding and NVIDIA for support.

REFERENCES

[1] V. Sze, M. Budagavi, G. J. Sullivan, and E. , *High Efficiency Video Coding: Algorithms and Architectures*. Springer, 07 2014.

[2] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen, "Modality-correlation-aware sparse representation for rgb-infrared object tracking," *Pattern Recognition Letters*, 2018.

[3] K. Souza, A. d. A. Araújo, Z. Patrocínio Jr, and S. Guimarães, "Graph-based hierarchical video segmentation based on a simple dissimilarity measure," *Pattern Recognition Letters*, vol. 47, pp. 85–92, 10 2014.

[4] R. Prates and W. R. Schwartz, "Kernel multiblock partial least squares for a scalable and multicamera person reidentification system," *Journal of Electronic Imaging*, vol. 27, no. 3, pp. 1–33, 2018.

[5] J. Almeida, J. A. dos Santos, B. Alberton, L. P. C. Morellato, and R. da S. Torres, "Phenological visual rhythms: Compact representations for fine-grained plant species identification," *Pattern Recognition Letters*, vol. 81, pp. 90–100, 2016.

[6] F. Kriegel, R. Köhler, J. Bayat-Sarmadi, S. Bayerl, A. E. Hauser, R. Niesner, A. Luch, and Z. Cseresnyés, "Cell shape characterization and classification with discrete fourier transforms and self-organizing maps," *Cytometry Part A*, vol. 93, 10 2017.

[7] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *International Conference on Pattern Recognition*, 2004, pp. 32–36.

[8] J. Hu, G.-S. Xia, F. Hu, and L. Zhang, "Dense v.s. sparse: A comparative study of sampling analysis in scene classification of high-resolution remote sensing imagery," *ArXiv e-prints*, 02 2015.

[9] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Computer Vision and Pattern Recognition*, June 2014, pp. 1725–1732.

[10] J. Wehrmann, G. S. Simões, R. C. Barros, and V. F. Cavalcante, "Adult content detection in videos with convolutional and recurrent neural networks," *Neurocomputing*, vol. 272, pp. 432–438, 2018.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision & Pattern Recognition*, jun 2008.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International journal of computer vision*, vol. 111, no. 1, pp. 98–136, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[16] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," *Computer Vision and Pattern Recognition*, 2009.

[17] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition*, jun 2009.

[18] D. Sad, V. Mota, L. Maciel, M. B. Vieira, and A. de Albuquerque Araújo, "A tensor motion descriptor based on multiple gradient estimators," in *SIBGRAPI*, aug 2013.

[19] V. Mota, J. Souza, A. de Albuquerque Araújo, and M. B. Vieira, "Combining orientation tensors for human action recognition," in *SIBGRAPI*, aug 2013.

[20] V. F. Mota, E. A. Perez, L. M. Maciel, M. B. Vieira, and P.-H. Gosselin, "A tensor motion descriptor based on histograms of gradients and optical flow," *Pattern Recognition Letters*, vol. 39, pp. 85–91, April 2014.

[21] E. A. Perez, V. F. Mota, L. M. Maciel, D. Sad, and M. B. Vieira, "Combining gradient histograms using orientation tensors for human action recognition," in *International Conference on Pattern Recognition*, 2012, pp. 3460–3463.

[22] V. F. Mota, E. A. Perez, M. B. Vieira, L. M. Maciel, F. Precioso, and P.-H. Gosselin, "A tensor based on optical flow for global description of motion in videos," in *SIBGRAPI*, august 2012, pp. 298–301.

[23] F. L. M. Oliveira, H. Maia, V. Mota, M. Vieira, and A. Araujo, "A variable size block matching based descriptor for human action recognition," *Journal of Communication and Information Systems*, vol. 30, no. 1, 2015.

[24] T. Kobayashi and N. Otsu, "Motion recognition using local auto-correlation of spacetime gradients," *Pattern Recognition Letters*, vol. 33, no. 9, pp. 1188 – 1195, 2012.

[25] M. Faraki, M. Palhang, and C. Sanderson, "Log-euclidean bag of words for human action recognition," in *IET Computer Vision (IET-CV)*, 2014.

[26] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition*, 2012, pp. 1234–1241.

[27] R. Minhas, A. Baradarani, S. Seifzadeh, and Q. M. Jonathan Wu, "Human action recognition using extreme learning machine based on visual vocabularies," *Neurocomputing*, vol. 73, no. 10-12, pp. 1906–1917, Jun. 2010.

[28] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, Mar. 2013.

[29] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[30] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition*, 2011, pp. 3361–3368.

[31] H. Wang, A. Kläser, C. Schmid, and L. Cheng-Lin, "Action Recognition by Dense Trajectories," in *Conference on Computer Vision and Pattern Recognition*, Colorado Springs, United States, Jun. 2011, pp. 3169–3176.

[32] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition*, 2010.

[33] L. Shao and R. Gao, "A wavelet based local descriptor for human action recognition," in *British Machine Vision Conference*, 2010, pp. 72.1–10, doi:10.5244/C.24.72.

[34] O. Kihl, D. Picard, and P.-H. Gosselin, "A unified formalism for video descriptor," in *International Conference on Image Processing*, 2013.

[35] M. Jain, H. Jégou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Computer Vision and Pattern Recognition*, Apr. 2013.

[36] E. Vig, M. Dorr, and D. D. Cox, "Saliency-based selection of sparse descriptors for action recognition," *International Conference on Image Processing*, pp. 1405–1408, 2012.

[37] T. Zhou, N. Li, X. Cheng, Q. Xu, L. Zhou, and Z. Wu, "Learning semantic context feature-tree for action recognition via nearest neighbor fusion," *Neurocomputing*, vol. 201, pp. 1–11, 2016.

[38] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *European Conference on Computer Vision*, 2014, pp. 581–595.

[39] A. Liu, Y. Su, W. Nie, and M. Kankanhalli, "Hierarchical clustering multi-task learning for joint human action grouping and recognition," *Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 102–114, Jan 2017.

[40] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *International Conference on Computer Vision*, 2011.

[41] S. Avila, N. Thome, M. Cord, E. Valle, and A. De Albuquerque AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[42] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, 2010.

[43] C. Caetano, S. Avila, W. R. Schwartz, S. J. F. G. aes, and A. de A. Araújo, "A mid-level video representation based on binary descriptors: A case study for pornography detection," *Neurocomputing*, vol. 213, pp. 102 – 114, 2016, binary Representation Learning in Computer Vision.

[44] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "Pornography classification: The hidden clues in video space-time," *Forensic Science International*, 2016.

[45] C. Decaestecker, O. Debeir, P. Van Ham, and R. Kiss, "Can anti-migratory drugs be screened in vitro? a review of 2d and 3d assays for the quantitative analysis of cell migration," *Medicinal Research Reviews*, vol. 27, no. 2, pp. 149–176, 2007.

[46] N. Ramnath and P. Creaven, "Matrix metalloproteinase inhibitors," *Current Oncology*, vol. 6, March 2004.

[47] V. F. Mota, H. Oliveira, S. Scalzo, D. D., R. J. Santos, J. A. dos Santos, and A. A. Araújo, "From video pornography to cancer cells: a tensor framework for spatiotemporal description," *Multimedia Tools and Applications. Under Review*, 2018.

[48] V. F. Mota, M. B. Vieira, and A. A. Araújo, "Busca por imagens e vídeos com base no conteúdo visual: Uma introdução," in *Anais da VII Escola Regional de Informática de Minas Gerais*, 2012, pp. 1–24.

[49] V. F. Mota, G. D. Dias, W. Santos, M. Vieira, and A. Araujo, "Tensor clustering for human action recognition," in *Workshop of Works in Progress (SIBGRAPI)*, 2015.

[50] H. A. Maia, A. M. O. Figueiredo, F. L. M. Oliveira, V. F. Mota, and M. B. Vieira, "A video tensor self-descriptor based on variable size block matching," *Journal of Mobile Multimedia*, vol. 11, pp. 90–102, 2015.

[51] A. M. O. Figueiredo, M. Caniato, V. F. Mota, R. L. S. Silva, and M. B. Bernardes, "A video self-descriptor based on sparse trajectory clustering," in *International Conference in Computer Science and its Applications*, 2016, pp. 571–583.

[52] C. S. Lenzoni, G. De Paula, L. W. De Freitas, V. F. Mota, L. Pires, and N. Fernandes, "Ferramenta de assistência médica para o estudo de declínio cognitivo em pacientes com doença renal crônica," in *Workshop of Works in Progress/XXIX Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2016, pp. 571–583.

[53] A. M. O. Figueiredo, H. A. Maia, F. L. M. Oliveira, V. Mota, and M. B. Vieira, "A video tensor self-descriptor based on block matching in: Computational science and its applications," in *International Conference in Computer Science and its Applications*, 2014, pp. 401–414.

[54] F. L. M. Oliveira, H. Maia, V. F. Mota, M. B. Vieira, and A. A. Araujo, "Video tensor self-descriptor based on variable size block matching," in *WVHAR - Workshop on Vision-based Human Activity Recognition (SIBGRAPI)*, 2014.

[55] C. E. Santor Jr, J. I. C. Souza, V. F. Mota, G. Sad, G. Gorgulho, and A. A. Araújo, "Panview: An extensible panoramic video viewer for the web," in *Latin American Web Congress (LAWEB)*, 2014.