DOMINANT SPATIO-TEMPORAL MODULATIONS AND ENERGY TRACKING IN VIDEOS: APPLICATION TO INTEREST POINT DETECTION FOR ACTION RECOGNITION

Christos Georgakis[1], Petros Maragos[1], Georgios Evangelopoulos[2] and Dimitrios Dimitriadis[3]

[1]School of E.C.E., National Technical University of Athens, Greece
ch.georgakis@gmail.com maragos@cs.ntua.gr
[2]Dept. Computer Science, University of Houston, TX, U.S.A. gevangel@central.uh.edu
[3]AT&T Labs-Research, 180 Park Ave, Florham Park, NJ, U.S.A. ddim@research.att.com

ABSTRACT

The presence of multiband amplitude and frequency modulations (AM-FM) in wideband signals, such as textured images or speech, has led to the development of efficient multicomponent modulation models for low-level image and sound analysis. Moreover, compact yet descriptive representations have emerged by tracking, through non-linear energy operators, the dominant model components across time, space or frequency. In this paper, we propose a generalization of such approaches in the 3D spatio-temporal domain and explore the potential of incorporating the Dominant Component Analysis scheme for interest point detection and human action recognition in videos. Within this framework, actions are implicitly considered as manifestations of spatio-temporal oscillations in the dynamic visual stream. Multiband filtering and energy operators are applied to track the source energy in both spatial and temporal frequency bands. A new measure for extracting keypoint locations is formulated as the temporal dominant energy computed over the spatial dominant components, in terms of their modulation energy, of input video frames. Theoretical formulation is supported by evaluation and comparisons in human action classification, which demonstrate the potential of the proposed spatio-temporal detector.

Index Terms— Human action recognition in videos, spatio-temporal interest point detectors, multiband filtering, multicomponent AM-FM models, dominant component analysis

1. INTRODUCTION

Local space-time features have gained growing popularity over the past years for the human action recognition task. Video representations in terms of such features exhibit efficiency in distinguishing among action classes, while bypassing the need for precise background subtraction or tracking. The *bag-of-features* approach [1, 2, 3], that constructs a histogram-based representation, starting from an orderless collection of spatio-temporal features, has proved successful in conjunction with SVM classification frameworks.

Feature extraction is usually accomplished in two discrete steps. First, spatio-temporal *interest point detectors* select highly informative and discriminative visual units that typically correspond to the local maxima of a proposed saliency measure. A list of popular detectors, extensively used within action recognition frameworks, includes Harris3D [4], Cuboid [5] and Hessian [6]. Next, *local descriptors* employ image measurements, such as gradients and optical flow, to encode local appearance and motion information in the space-time neighborhood of the detected points. Among the most prevalent descriptors are Cuboid [5], HOG/HOF [2], HOG3D [7] and Extended SURF [6].

Interest points in space and time sought by detectors, correspond in most cases to spatially prominent locations characterized by sharp changes in the direction of motion. The extension of Harris detector to 3D by Laptev and Lindeberg [4] was one of the first attempts to address keypoint extraction in videos, and has been deemed a benchmark for quantitative evaluation of other detectors in publicly available datasets. Dollár *et al*. [5] argued against detections on the basis of spatio-temporal "cornerness", and introduced a response function which relies on spatial Gaussian smoothing and a quadrature pair of Gabor filters applied in the temporal dimension. Although the detector favors periodic movements, these are intrinsically related to the single center frequency used to tune the temporal filters, thus ignoring periodicity content present within other bands. Recently, Bregonzio *et al*. [8] approached feature extraction through spatial Gabor filters at multiple orientations on frame-difference based regions of interest. Willems *et al*. [6] formulate a saliency measure as the determinant of the spatio-temporal Hessian matrix, which also serves for scale selection. The efficiency of the aforementioned detectors (Cuboid, Harris3D and Hessian) was systematically evaluated in realistic video settings by Wang *et al*. [3]. Notably, dense sampling at a regular spatio-temporal grid was shown to be superior for recognition on two action datasets, Hollywood2 and UCF.

In this paper, we apply energy tracking for spatio-temporal interest point detection in videos, based on unveiling locally dominant modulation structures in both space and time. Modulation components in wideband signals have been previously applied with success in nonlinear speech analysis [9, 10, 11], texture analysis and image segmentation [12, 13, 14]. Herein, we propose a novel, to the best of our knowledge, generalization of multicomponent AM-FM models in the spatio-temporal domain for video analysis and employ the dominant energy volume for action recognition based on sparse, non-regular feature sampling. More specifically, we combine multiband filtering and energy tracking for dominant modulation component representations; these are sequentially applied first in the spatial and following in the temporal domain. The resulting response function can be viewed conceptually as a saliency measure in terms of the dominant spatio-temporal modulation energy. An illustration of the proposed detector framework is given in Figure 1.

The remainder of this paper is organized as follows. In Section 2, we provide background on energy operators and modulation energy tracking along with an outline of the Dominant Component Analysis scheme. Section 3 formulates the proposed spatio-temporal

This work was done when all authors were at NTUA. It was partially supported by the EU under the research projects DictaSign with grant FP7-ICT-3-231135 and DIRHA with grant FP7-ICT-7-288121.

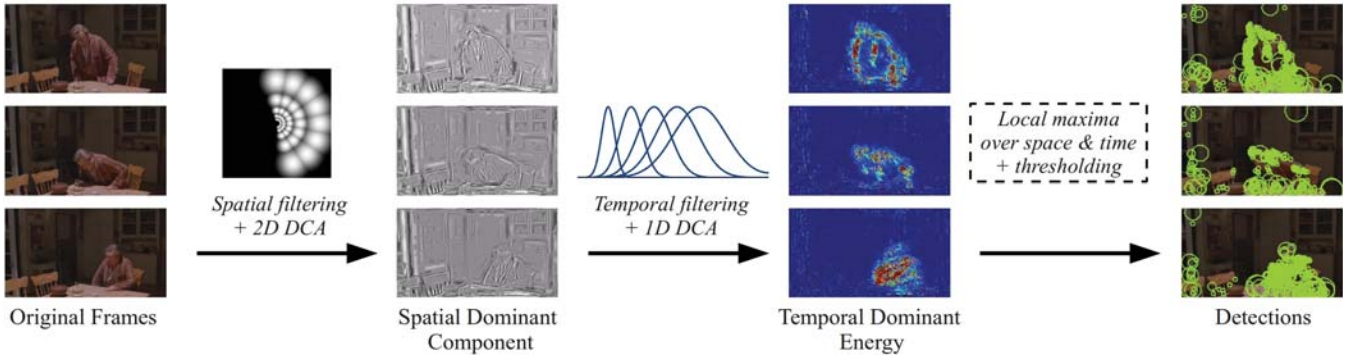

Fig. 1: Spatio-temporal interest point detection stages (better viewed in color) on three sample frames from *SitDown* action of Hollywood2 Actions Dataset. Original frames (left), converted to grayscale, undergo spatial Gabor filtering (five scales and eight orientations) and the Spatial Dominant Component (max-energy filter output) is extracted pointwise (left middle) using 2D Dominant Component Analysis (DCA). Next, temporal Gabor filtering (five frequencies) and 1D DCA yield the Temporal Dominant Energy at each point (right middle). Final detections (right) are the prominent local maxima on the 3D energy volume.

dominant modulation energy, both conceptually and algorithmically. In Section 4, we present the experimental framework and action recognition results. Finally, Section 5 concludes the paper.

2. ENERGY OPERATORS AND MULTIBAND MODULATION ENERGY TRACKING

The development of the nonlinear, differential *Teager-Kaiser Energy Operator (TKEO)* [9] $\Psi[s(t)] \triangleq [s'(t)]^2 - s(t)s''(t)$ has facilitated the energy representation and demodulation of signals modeled by non-stationary sinusoids, with amplitude and frequency modulation (AM-FM), of the form $s(t) = \alpha(t)\cos(\phi(t))$. Under realistic conditions [10], when applied to such a signal the operator Ψ yields as output $\Psi[s(t)] \approx [\alpha(t)\phi'(t)]^2$, which equals to the energy of the oscillatory source. Maragos and Bovik [12] extended the energy operator to signals of higher dimensions, rendering it applicable to real-valued grayscale images $I(x, y)$ as a 2D operator

$$\Phi(I(x,y)) \triangleq \|\nabla I(x,y)\|^2 - I(x,y)\nabla^2 I(x,y). \quad (1)$$

However, a typically wideband speech or image signal (*e.g.* texture) entails the coexistence of multiple modulations in the spectrum. To cope with this, speech can be represented through a multicomponent AM-FM model as a sum of non-stationary, narrowband sinusoids [10]. Estimation of each component's Teager-Kaiser energy presupposes a decomposition on different frequency bands. An efficient solution to the component separation and energy tracking problem, resulting also in noise suppression, is through bandpass Gabor filtering [10] and the *Regularized* or *Gabor TKEO* [11] operator

$$\Psi_h(s(t) * h(t)) = [s(t) * h'(t)]^2 - [s(t) * h(t)][s(t) * h''(t)], \quad (2)$$

where $h(t)$ is the filter's impulse response.

Similarly to the 1D case, textured or wideband images in general, characterized by non-smooth structures, are more accurately represented by a multicomponent AM-FM image model

$$I(x,y) = \sum_{k=1}^{K} \alpha_k(x,y)\cos(\phi_k(x,y)) \quad (3)$$

instead of a single modulation component [13]. In this decomposition, for each of the K narrowband components the amplitude modulating signal $\alpha_k(x, y)$ conveys local image contrast information, while local image structure properties, such as scale and orientation, can be disclosed by the frequency modulation vector $\boldsymbol{\omega}_k(x, y) = \nabla\phi_k(x, y)$. The output of the 2D *TKEO* (1) on the k-th narrowband component approximates the energy product of the squared amplitude and frequency magnitude $\Psi(I_k) \approx {a_k}^2 ||\boldsymbol{\omega}_k||^2$. Kokkinos *et al*. [14] coupled multiband filtering and modulation energy estimation by introducing the 2D counterpart of the *Gabor TKEO* (2)

$$\Phi_g(I * g) = \|I * \nabla g\|^2 - (I * g)\left(I * \nabla^2 g\right), \quad (4)$$

where $g(x, y)$ denotes the impulse response of a real-valued Gabor filter. The wideband image is decomposed into a fixed component tessellation, based on the pre-defined 2D Gabor filter configuration, with the oscillation energy of each local modulation continuously estimated across image locations.

Analysis schemes procured by multiband filtering and component demodulation have shown efficacy in reconstructing and uncovering the most significant structures of textured images [13, 14]. The *Dominant Component Analysis (DCA)* method [13] seeks at each pixel the channel associated to the component whose response prevails in the local image spectrum. In [14] the modulation energy was proposed as an alternative to the amplitude estimate, as the criterion for dominant channel selection. Given that modulation energy jointly encompasses amplitude and frequency content, the energy-based DCA scheme attributes saliency, in locally high-contrast regions or low-contrast but high modulating frequency magnitude.

3. SPATIO-TEMPORAL DOMINANT ENERGY AND KEYPOINT DETECTION

As a human action evolves over time, image intensities within the regions containing significant body parts movements vary with time in a way that intuitively resembles the behavior of 1D time-domain AM-FM signals. Thus, such quasi-periodic limb motions occurring at different speeds in the scene are highly likely to correspond to sinusoidal components of non-stationary temporal amplitude and frequency modulations. Motivated by such analogies between space-time activity and temporal modulations and the representational potential of energy-based DCA, we propose a 3D generalization of the model and apply it for keypoint detection in videos of human actions.

The framework involves extraction of the temporal DCA on the dominant components of the video sequence frames. Initially, 2D

energy-based DCA is applied on each grayscale-converted frame to emphasize the prominent texture variations and meaningful object boundary information. For spatial multiband filtering, a set of $K = 40$ isotropic, complex Gabor filters are arranged in five scales and eight orientations to densely cover the 2D frequency plane. The energy of the narrowband, filtered components per frame is computed using the 2D Gabor TKEO (4). According to the max-energy DCA criterion, the dominant channel $i(x, y)$ at each spatial location (x, y) is the one yielding the maximum value in the operator output

$$i(x, y) = \underset{1 \le k \le K}{\arg\max} \{\Phi_g(I * g_k)(x, y)\} . \quad (5)$$

Subsequently, each location is represented only by $(I*g_i)(x, y)$, *i.e.*, the i-th filtered bandpass image value at pixel (x, y). These DCA-synthesized filtered images for all frames form a new 3D volume, henceforth termed *Spatial Dominant Component (SDC)*.

To decouple temporal modulation components, multiband temporal filtering is applied on volume SDC (x, y, t) by five 0.75-bandwidth temporal Gabor filters, linearly spaced to span the normalized frequency interval $[0 - \text{framerate}/2]$ Hz. Similar to the spatial processing step, the 1D Gabor TKEO (2) accounts for both tasks of filtering and modulation energy tracking for each of the $L = 5$ channels. The 1D energy-based DCA methodology is applied to select voxelwise the dominant temporal channel

$$j(x, y, t) = \underset{1 \le l \le L}{\arg\max} \{\Psi_h (\text{SDC}(x, y, t) * h_l(t))\} , \quad (6)$$

where h_l the impulse response of the l-th filter.

The proposed spatio-temporal energy function is established as the *Temporal Dominant Energy of the Spatial Dominant Component*

$$R(x, y, t) = \Psi_h (\text{SDC}(x, y, t) * h_j(t)) , \quad (7)$$

where $j(x, y, t)$ is given by (6). A set $\mathcal{M}$ of interest points is comprised by detecting the local maxima over space and time of the volume energy function (7). False alarms are banished via the imposition of an appropriate global threshold T to the local maxima values, leading to the final detection subset $\mathcal{D} \subseteq \mathcal{M}$

$$\mathcal{D} = \left\{(x, y, t) \in \mathcal{M} : R(x, y, t) > T \cdot \max_{\mathcal{M}}(R)\right\} . \quad (8)$$

4. EXPERIMENTAL RESULTS

The novel interest point detector, called DCA3D, is tested on four action classes of the challenging *Hollywood2 Actions Dataset* [15], namely *AnswerPhone, SitDown, StandUp* and *FightPerson*. The global threshold was set to $T = 0.07$, after experimenting with several values, by optimizing keypoint density, according to recognition results. To generate feature vectors we choose the HOG/HOF descriptor in its multiscale implementation, where the detection scales (σ, τ) were set equal to the standard deviation of the dominant filter Gaussian envelope in each respective dimension and then quantized to the default values in [2]. With all samples processed at half spatial resolution, the 3D patch size was defined by $\Delta_x = \Delta_y = 10\sigma, \Delta_t = 8\tau$. For the rest of the evaluation framework, we followed a *bag-of-features* SVM approach identical to [3], using the standard Gaussian kernel. As a baseline, we use the combination of Harris3D detector [4] and the standard multiscale HOG/HOF descriptor [2], as an optimal, state-of-the-art performance method [3].

Recognition scores, expressed as Average Precision for each class, are reported in Table 1 for the evaluated detection algorithms.

Actions	DCA3D	Harris3D
AnswerPhone	25.8%	29.0%
SitDown	61.4%	58.8%
StandUp	72.2%	68.2%
FightPerson	83.3%	94.7%

Table 1: Average Precision (AP) on four actions of Hollywood2 Dataset for HOG/HOF features using DCA3D and Harris3D.

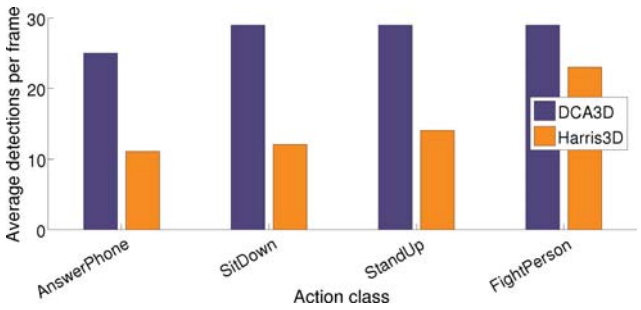

Fig. 2: Average number of detected points per frame computed over all samples belonging to each action class, for both detectors.

Results with DCA3D compare favorably to those obtained from Harris3D for two action classes, *StandUp* and *SitDown*, while performing less accurately on class labeling for the remaining two. Average detections per frame, extracted for samples with the same action by each detector, are illustrated in Figure 2. As can be seen, there is a certain discrepancy in feature density in most action classes between the two methods, with DCA3D yielding about 1.75 times the Harris3D detections in average. This behavior is partially affected by the empirical selection of the threshold T which can be optimally tuned through systematic validation techniques, such as ROC performance analysis. In addition, feature density distribution is more uniform across classes for DCA3D and thus less dependent on action type, which is rather undesirable for this task. A remedy for that could lie on action-specific learning of the global threshold in order to derive a representation with class-adaptive feature sparsity. The relatively low performance on *FightPerson* can be attributed to camera motion and fast shot changes, typical for such action samples, on which the detector is more susceptible due to its high-response to multiscale motion. Their effects could be alleviated using camera motion tracking and stabilization or, alternatively, shot-based keypoint extraction, after shot-detection on the action sequence.

To demonstrate competing performance on a simpler benchmark database, we applied our 3D detector framework on the KTH database and evaluation framework [2, 3]. We restricted to using only HOF as features, as state-of-the-art average accuracy of 92.1% has been obtained using Harris3D/HOF [3]. The detection threshold was empirically set to a fixed value $T = 0.14$, by imposing an

	Walking	Jogging	Running	Boxing	Waving	Clapping
Walking	1.00	0.00	0.00	0.00	0.00	0.00
Jogging	0.47	0.31	0.21	0.01	0.00	0.00
Running	0.23	0.17	0.59	0.01	0.00	0.00
Boxing	0.01	0.00	0.00	0.92	0.01	0.06
Waving	0.00	0.00	0.00	0.00	0.92	0.08
Clapping	0.00	0.00	0.00	0.01	0.00	0.99

Table 2: Confusion matrix for DCA3D/HOF on KTH Dataset.

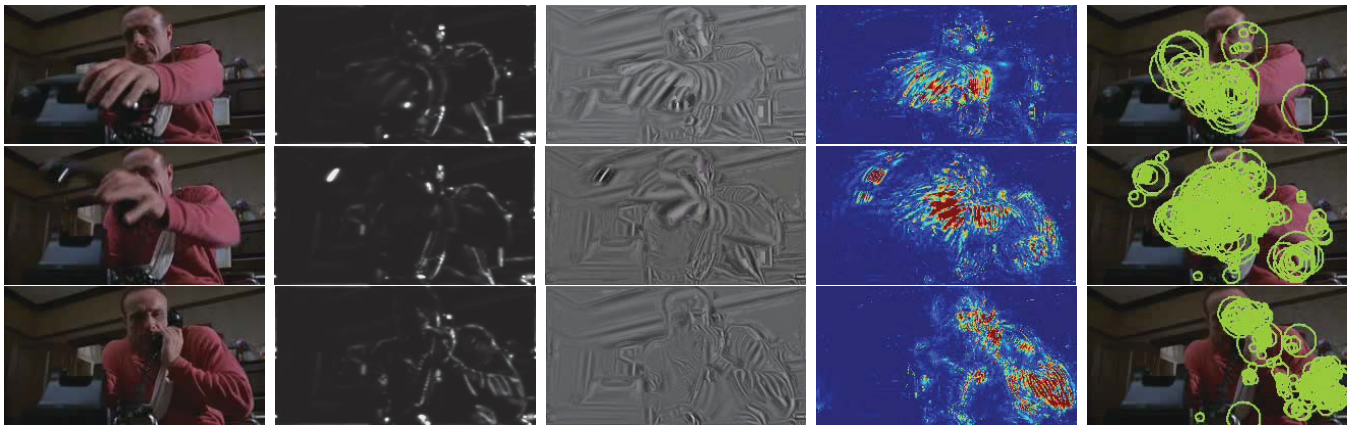

Fig. 3: DCA3D detector on a sample from *AnswerPhone* action of the Hollywood2 Dataset. Intermediate outputs and detections are shown on three action instances corresponding roughly to its starting, middle and ending point. Note how the Spatial Dominant Component (and Energy) de-emphasizes low-energy textured areas, while preserving main image structure, like edges and boundaries. The Temporal Dominant Energy assigns salient values to regions of complex motion, *e.g.* person's hand or elbow, resulting on a dense concentration of detected points.

average rate of 10 detections/frame across train and test sets. The confusion matrix for the six action classes can be seen on Table 2, corresponding to an average DCA3D/HOF accuracy of 78.8%. It is noteworthy that, without parameter validation and tuning, DCA3D surpasses 90% in four classes and achieves superior performance in three (*Walking, Waving, Clapping*) compared to the state-of-the-art Harris3D/HOF [2]. The misclassifications for *Jogging* and *Running* actions can be suppressed by using class-based detected density and validation-based parameter learning.

5. CONCLUSION

We proposed a new video analysis method that builds on both texture and motion decomposition to detect and track multiband spatio-temporal modulation components. Their energy has been formulated as the basis of a spatio-temporal interest point detector for sparse feature extraction from video volumes and applied for recognizing actions in movie clips. The proposed representation is effective in capturing local oscillations in the spatial domain, as non-stationary contrast, scale and orientation changes, and the temporal, as varying, non-stationary movements of inherent periodicity. Preliminary experimental results show comparable performance to an efficient, state-of-the-art detector on actions from a challenging database (Hollywood2). Our aim is to refine the tracking process and parameter selection, investigate the potential for action-driven feature selection, and extend our experimental comparisons to more action datasets.

6. REFERENCES

[1] J.C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *Intl. Journ. Comp. Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[2] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. CVPR*, 2008.

[3] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. BMVC*, 2009.

[4] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. ICCV*, 2003.

[5] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE Intl. Workshop VSPETS,*, 2005, pp. 65–72.

[6] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. ECCV*, 2008.

[7] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. BMVC*, 2008.

[8] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Proc. CVPR*, 2009.

[9] J.F. Kaiser, "On a simple algorithm to calculate the energy of a signal," in *Proc. ICASSP*, 1990.

[10] P. Maragos, J.F. Kaiser, and T.F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Trans. Sig. Proc.*, vol. 41, no. 10, pp. 3024–3051, 1993.

[11] D. Dimitriadis and P. Maragos, "Continuous energy demodulation methods and application to speech analysis," *Speech Communication*, vol. 48, no. 7, pp. 819–837, 2006.

[12] P. Maragos and A.C. Bovik, "Image demodulation using multidimensional energy separation," *J. Opt. Soc. Amer. A*, vol. 12, no. 9, pp. 1867–1876, 1995.

[13] J.P. Havlicek, D.S. Harding, and A.C. Bovik, "Multidimensional quasi-eigenfunction approximations and multicomponent am-fm models," *IEEE Trans. Im. Proc.*, vol. 9, no. 2, pp. 227–242, 2000.

[14] I. Kokkinos, G. Evangelopoulos, and P. Maragos, "Texture analysis and segmentation using modulation features, generative models, and weighted curve evolution," *IEEE Trans. PAMI*, vol. 31, no. 1, pp. 142–157, 2009.

[15] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. CVPR*, 2009.