In: *Proc. SSVM 2017: Scale-Space and Variational Methods in Computer Vision*, Kolding, Denmark, Springer LNCS vol 10302, pages 16–28.

Dynamic texture recognition using time-causal spatio-temporal scale-space filters*

Ylva Jansson and Tony Lindeberg

Computational Brain Science Lab
Department of Computational Science and Technology
School of Computer Science and Communication
KTH Royal Institute of Technology, Stockholm, Sweden

Abstract. This work presents an evaluation of using time-causal scale-space filters as primitives for video analysis. For this purpose, we present a new family of video descriptors based on regional statistics of spatio-temporal scale-space filter responses and evaluate this approach on the problem of dynamic texture recognition. Our approach generalises a previously used method, based on joint histograms of receptive field responses, from the spatial to the spatio-temporal domain. We evaluate one member in this family, constituting a joint binary histogram, on two widely used dynamic texture databases. The experimental evaluation shows competitive performance compared to previous methods for dynamic texture recognition, especially on the more complex DynTex database. These results support the descriptive power of time-causal spatio-temporal scale-space filters as primitives for video analysis.

1 Introduction

The ability to derive properties of the surrounding world from time-dependent visual input is a key functionality of a computer vision system, and necessary for any artificial or biological agent that is to use visual input for interpreting a dynamic environment. For this purpose, there has been intensive research into areas such as action recognition, dynamic texture and scene understanding, automatic surveillance, video-indexing and retrieval, etc.

For biological vision, local image measurements in terms of receptive fields constitute the first processing layers [1]. In computer vision, spatial receptive fields based on the Gaussian scale-space concept have been demonstrated to be a powerful front-end for solving a large range of visual tasks. The theoretical

* The support from the Swedish Research Council (Contract 2014-4083) and Stiftelsen Olle Engkvist Byggmästare (Contract 2015/465) is gratefully acknowledged.

properties of scale-space filters enable the design of methods invariant or robust to natural image transformations [2–4]. Also, such axiomatically derived first processing layers, which can be shared among different tasks, free resources both for learning higher level features from data and during on-line processing. This could prove especially useful for high-dimensional video data.

For a real-time visual system or to model biological vision, such a visual front-end cannot utilise information from the future. For time-critical applications (such as self-driving cars) where also a small difference in response time matters, the ad-hoc solution of using a time-delayed truncated Gaussian temporal kernel would imply unnecessarily long temporal delays. Recently, a new framework for *time-causal spatio-temporal scale-space filters*, or equivalently *spatio-temporal receptive fields*, was introduced by Lindeberg [2]. These idealised receptive fields show a strong connection to biology in the sense that they very well model receptive field shapes of neurons in the LGN and V1 [2, 4]. The purpose of this study is a first evaluation of using these time-causal spatio-temporal receptive fields as visual primitives for video analysis.

As a first application, we have chosen the problem of dynamic texture recognition. A dynamic texture or spatio-temporal texture can be naively defined as "texture + motion" or more formally as a spatio-temporal pattern that exhibits certain stationarity properties and self-similarity over both space and time [5]. Examples of dynamic textures are windblown vegetation, fire, waves, a flock of flying birds or a flag flapping in the wind. Recognising different types of dynamic textures is important for visual tasks such as automatic surveillance (e.g. detecting forest fires), video indexing and retrieval (e.g. return all images set on the sea) and to enable artificial agents to understand and interact with the world.

In this paper, we start by presenting a new family of video descriptors in the form of joint histograms of spatio-temporal receptive field responses, thereby generalising a previous method by Linde and Lindeberg [6] from the spatial to the spatio-temporal domain. We then evaluate one member of this family constituting a joint binary histogram on two widely used dynamic texture databases. It will be shown that our preliminary descriptor shows highly competitive performance compared to previous methods for dynamic texture recognition, thus supporting the applicability of these time-causal spatio-temporal receptive fields as primitives for video analysis.

2 Related work

Some of the first methods for dynamic texture recognition were based on *optic flow*, see e.g. Nelson and Polana [7]. Another early approach for both synthesis and recognition was to model dynamic textures as *linear dynamical systems* (LDS), see e.g. work by Soatto et al. [8]. To enable recognition less dependent on global spatial appearance, the LDS approach has been extended to bags of dynamical systems (BoS) as by Ravichandran et al. [9] and Wang et al. [10], where the latter approach also combines local LDS descriptors with soft coding and an extreme learning machine (ELM) classifier. Previous approaches utilising

spatio-temporal filtering are e.g. the oriented energy representations by Wildes and Bergen [11] and Derpanis and Wildes [12], where the latter represents *pure dynamics* of spatio-temporal textures by capturing space-time orientation by means of 3D Gaussian derivative filters. Gonçalves et al. [13] instead jointly model appearance and dynamics using spatio-temporal Gabor filters with different preferred spatial orientations and speeds. Neither of these approaches utilise joint statistics of filter responses.

Methods that model *local space-time structure* of dynamic textures are e.g. local binary patterns (LBP) (Zhao et al. [14]) that capture the joint binarised distribution of local neighbourhoods of pixels, either for 3D space-time volumes (VLBP) or on three orthogonal planes (LBP-TOP), where the latter reduces the computational load by considering the XY, YT, and XT planes separately. Extensions to LBP-TOP are e.g. utilising averaging and principal histogram analysis to get more reliable statistics (Ren et al. [15]) or multi-clustering of salient features to identify and remove outlier frames (AFS-TOP) (Hong et al. [16]). A related approach is multi-scale binarised statistical image features (MBSIF-TOP) introduced by Arashloo and Kittler [17], which capture local image statistics by means of *filters learned from data* by independent component analysis. Tensor dictionary learning (OTD) (Qu et al. [18]) is instead a sparse coding based approach for learning a dictionary for local space-time structure. Previous approaches using non-binary *joint histograms* for image analysis include Schiele and Crowley [19] and Linde and Lindeberg [6], but many later methods have often used either marginal histograms or relative feature strength to capture image statistics. Xu et al. [20] utilise the self-similarity of dynamic textures by creating a descriptor from the fractal dimension of motion features (DFS) and Ji et al. [21] present a method based on wavelet domain fractal analysis (WMFS).

There are also approaches combining several different descriptors such as DL-PEGASOS by Ghanem and Ahuja [22] that uses LBP, PHOG and LDS descriptors together with maximum margin distance learning or Yang et al. [23] using ensemble SVMs to combine LBP, shape-invariant co-occurrence patterns (SCOPs) and chromatic information with dynamic information represented by LDS. Qi et al. [24] present a dynamic texture descriptor leveraging deep learning to transfer prior knowledge from the image domain by extracting global features using a pretrained convolutional neural network. Compared to the time-causal scale-time kernel proposed by Koenderink [25] it should be noted that the time-causal limit kernel used in this paper is *time-recursive*, whereas no time-recursive formulation is known for the scale-time kernel.

3 Spatio-temporal receptive field model

The spatio-temporal scale-space framework and receptive field model used in this work is that of Lindeberg [2]. The axiomatically derived scale-space kernel for spatial scale s and temporal scale τ is of the form

$$T(x, y, t;\ s, \tau, u, v, \Sigma) = g(x - ut, y - vt;\ s, \Sigma)\, h(t;\ \tau) \tag{1}$$

where (x, y) denotes the image coordinates; t denotes time; $h(t;\ \tau)$ denotes a temporal smoothing kernel and $g(x - ut, y - vt;\ s, \Sigma)$ denotes a spatial affine Gaussian kernel with spatial covariance matrix Σ that moves with image velocity (u, v). Here, we restrict ourselves to rotationally symmetric Gaussian kernels over the spatial domain and to smoothing kernels with zero image velocity leading to space-time separable receptive fields. The temporal smoothing kernel $h(t;\ \tau)$ used here is the time-causal kernel composed from coupling truncated exponential functions in cascade, with a composed scale-invariant limit kernel having a Fourier transform of the form [2, Eq. 38]

$$\hat{\Psi}(\omega;\ \tau, c) = \prod_{k=1}^{\infty} \frac{1}{1 + i\, c^{-k} \sqrt{c^2 - 1} \sqrt{\tau}\, \omega} \tag{2}$$

where $c > 1$ is the distribution parameter for the logarithmic distribution of intermediate scale levels. For practical purposes, the limit kernel is approximated by a finite number, K, of recursive filters coupled in cascade according to [2, Section 6]. We here use $c = 2$ and $K \geq 7$. The time-recursive formulation means there is no need for saving a temporal buffer of previous frames — computing the scale-space representation for a new frame only requires information from the present moment and the scale-space representation for the preceding frame. The *spatio-temporal receptive fields* are in turn defined as partial derivatives of the spatio-temporal scale-space representation of a video $f(x, y, t)$

$$L_{x^{m_1} y^{m_2} t^n}(\cdot, \cdot, \cdot; s, \tau, u, v, \Sigma) = \partial_{x^{m_1} y^{m_2} t^n}(T(\cdot, \cdot, \cdot;\ s, \tau, u, v, \Sigma) * f(\cdot, \cdot, \cdot)) \tag{3}$$

leading to a spatio-temporal *N-jet* representation of local space-time structure

$$\{L_x, L_y, L_t, L_{xx}, L_{xy}, L_{yy}, L_{xt}, L_{yt}, L_{tt}, ...\}. \tag{4}$$

We also perform scale normalisation of partial derivatives as described in [2]. A subset of receptive fields/scale-space derivative kernels can be seen in Figure 1. For details concerning the spatio-temporal scale-space representation and the discrete implementation, we refer to [2].

4 Video descriptors

We here describe our proposed family of video descriptors. The descriptor is computed in three main steps: (i) computation of local spatio-temporal receptive field responses, (ii) dimensionality reduction with PCA and (iii) aggregating joint statistics of receptive field responses into a multidimensional histogram.

4.1 Receptive field responses

The first processing step is to compute spatio-temporal receptive field responses $F = [F_1, F_2, ...F_N]$ over all individual pixels (x, y, t) in a space-time region for N scale-space derivative filters. These could include a range of different spatial

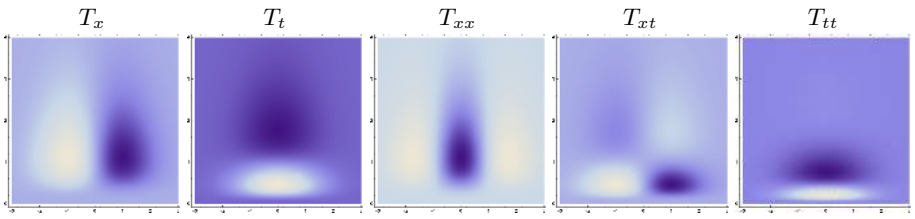

Fig. 1. Time-causal spatio-temporal scale-space derivative kernels $T_{x^m t^n}(x, t;\ s, \tau)$, with a Gaussian kernel over space and the time-causal limit kernel [2] over time, shown over a 1+1-D space-time for the space-time separable case when $v = 0$. ($s = 1$, $\tau = 1$, $K = 8$, $c = 2$) (Horizontal axis: space, $x \in [-3, 3]$. Vertical axis: time, $t \in [0, 3]$)

and temporal scales to enable capturing image structures of different spatial extent and temporal duration. Computations are separable in all dimensions and performed frame by frame, utilising recursive smoothing along the temporal dimension. In contrast to previous methods utilising spatio-temporal filtering, such as [12, 13], our method includes a diverse group of partial derivatives from the spatio-temporal N-jet as opposed to a single filter type.

4.2 Dimensionality reduction with PCA

When combining a large number of local image measurements, not all dimensions will carry meaningful information. For this reason, we perform dimensionality reduction with PCA of the local receptive field responses, as was empirically shown to give good results for spatial images in [6], resulting in a local feature vector $\tilde{F}(x, y, t) = [\tilde{F}_1, \tilde{F}_2, ...\tilde{F}_M] \in \mathbb{R}^M$ $M \leq N$. The number of components M can be adapted to requirements for descriptor size and need for detail in modelling the local image structure. Dimensionality reduction can also be skipped if working with a smaller number of receptive fields.

4.3 Joint receptive field histograms

When creating the joint histogram of receptive field responses, each feature dimension is partitioned into r number of equidistant bins in the range $[\text{mean}(\tilde{F}_i) - d\ \text{std}(\tilde{F}_i),\ \ \text{mean}(\tilde{F}_i) + d\,\text{std}(\tilde{F}_i)]$. This gives $n_{cells} = r^N$ distinct histogram bins. Such a joint histogram of spatio-temporal filter responses explicitly models the co-variation of different types of image measurements, in contrast to descriptors based on marginal distributions or relative feature strength. Each histogram cell will correspond to a certain "template" local space-time structure, similar to e.g. VLBP [14] but notably represented and computed using different primitives. The histogram descriptor thus captures the frequency of such local space-time structures in a video and the number of different "templates" will be decided by the number of receptive fields/principal components and the number of bins.

If represented naively, a joint histogram could imply a prohibitively large descriptor. However, in practise the number of *non-zero* bins can be considerably lower than the maximum number of bins, which enables utilising a computationally efficient sparse representation as outlined in [6]. Also note that although we in this work aggregate statistics over entire videos, it is straight-forward to instead compute descriptors regionally over both time and space; and thus to classify new videos after having seen only a limited number of frames.

4.4 Binary histograms

When choosing $r = 2$ bins equivalent to a joint binary histogram, the local image structure is described by only the sign of the different image measurements [6]. This will make the descriptor invariant to uniform rescalings of the intensity values. Another attractive feature of the binary histogram is that a larger number of image measurements can be combined while still keeping down the descriptor dimensionality. Binary histograms have been proven an effective approach for other dynamic texture methods such as LBP and MBSIF-TOP. This is the descriptor version that we have chosen to investigate in this paper.

4.5 Choice of receptive fields and parameters

Varying the choice of receptive fields and method parameters, gives a family of different video descriptors. In this paper, we evaluate a single member of this family: A multi-scale representation based on the set of receptive fields

$$\{L_t, L_{tt}, L_x, L_y, L_{xx}, L_{yy}, L_{xy}, L_{xt}, L_{yt}, L_{xxt}, L_{yyt}, L_{xyt}\} \tag{5}$$

with $M = 15$ principal components and $r = 2$ number of bins with $\mathrm{d} = 5$. For the UCLA database, the Cartesian product of spatial scales (i.e. standard deviation for the scale-space kernel) $\sigma_s \in \{1, 2\}$ pixels and temporal scales $\sigma_\tau \in \{0.05, 0.1\}$ seconds were used, while for DynTex that has considerably higher spatial resolution, we instead used $\sigma_s \in \{2, 4\}$ pixels and $\sigma_\tau \in \{0.2, 0.4\}$ seconds.

5 Datasets

We evaluate our proposed method on several dynamic texture recognition/ classification benchmarks from two widely used dynamic texture databases: UCLA and DynTex. Sample frames from these databases are shown in Figure 2.

5.1 UCLA

The UCLA database was introduced by Soatto et al. [8] and is composed of 200 videos (160×110 pixels, 15 fps) featuring 50 different dynamic textures with 4 samples from each texture. The **UCLA50** benchmark [8] divides the 200 videos into 50 classes with one class per individual texture/scene. It should be noted

Fig. 2. Top row: Sample frames from the UCLA database. From left to right: “fire”, “sea”, “smoke” “waterfall”, “fountain”. Bottom row: Sample frames from the DynTex database. From left to right: ”flags”, ”sea”, ”traffic”, ”escalator”, ”fountain”.

that this partitioning is not *conceptual* in the sense of classes constituting different types of textures such as “fountains”, “sea” or “flowers” but instead targets instance specific (i.e. different fountains should be separated from each other) and viewpoint specific recognition.

Since for many applications it is more relevant to recognise different dynamic texture categories, a partitioning of the UCLA dataset into conceptual classes, **UCLA9**, was introduced by Ravichandran et al. [9] with the following classes: boiling water (8), fire (8), flowers (12), fountains (20), plants (108), sea (12), smoke (4), water (12) and waterfall (16). Because of the large overrepresentation of plant videos, in the **UCLA8** benchmark those are excluded to give a less misbalanced dataset, resulting in 92 videos from 8 conceptual classes.

5.2 DynTex

A larger and more diverse dynamic texture database, **DynTex**, was introduced by Péteri et al. [26], featuring a larger variation of dynamic texture types recorded under more diverse conditions (720×576 pixels, 25 fps). From this database, three gradually larger and more challenging benchmarks have been compiled by Dubois et al. [27]. The **Alpha** benchmark includes 60 dynamic texture videos from three different classes: sea, grass and trees. There are 20 examples of each class and some variations in scale and viewpoint. The **Beta** benchmark includes 162 dynamic texture videos from ten classes: sea, vegetation, trees, flags, calm water, fountain, smoke, escalator, traffic and rotation. There are 7 to 20 examples of each class. The **Gamma** benchmark includes 264 dynamic texture videos from ten classes: flowers, sea, trees without foliage, dense foliage, escalator, calm water, flags, grass, traffic and fountains. There are 7 to 38 examples of each class and this dataset has the largest intraclass variability in terms of scale, orientation, etc.

	UCLA8		**UCLA9**		**UCLA50**	
	SVM	NN	SVM	NN	SVM	NN
Ensemble SVMs [23]	-	-	-	-	**100.0**	-
MBSIF-TOP [17]	-	**97.8**	-	**98.8**	-	99.5
OTD [18]	**99.5**	97.0	98.2	97.5	99.8	98.5
Enhanced LBP [15]	-	-	-	98.2	-	**100.0**
Our approach	97.8	97.5	**98.6**	98.3	98.5	97.0
DFS [20]	99.2	-	97.5		89.5	**100.0**
WMFS [21]	97.0	97.2	97.1	97.0	99.8	99.1
DL-PEGASOS [22]	-	-	-	95.6	-	99.0
Oriented energy rep. [12]	-	-	-	-	-	81.0

Table 1. Comparison to state-of-the-art for the UCLA benchmarks.

6 Experiments

We present results both using a support vector machine (SVM) classifier and a nearest neighbour (NN) classifier, the latter to evaluate the performance also without hidden tunable parameters. For NN we use the χ^2-distance $d(x, y) = \sum_i (x_i - y_i)^2/(x_i + y_i)$ and for SVM a χ^2-kernel $e^{-\gamma d(x,y)}$. The same set of receptive fields and the same parameters (see Section 4) are used for all experiments and no extensive parameter tuning has been performed. For the UCLA benchmarks, we also use the non-cropped videos of size 160×110, instead of the most common setup which is to use manually extracted patches; thus our setup could be considered a slightly harder problem.

6.1 Experimental setup and results UCLA50

The standard test setup for the UCLA50 benchmark, which we adopt also here, is 4 fold crossvalidation where for each partitioning one of the four videos of each dynamic texture are held out for testing and the other three are used for training [8]. Test results are seen in Table 1. It can be seen that we achieve competitive results on this benchmark with three misclassified samples out of 200, giving a classification accuracy of 98.5 % using an SVM classifier. Here, the best results achieved by DFS, LBP and Ensemble SVMs reach 100 %. Inspecting the misclassified samples, we note that two of those are different plants from the same viewpoint being mixed up and the third one is an instance of a specific plant being misclassified as the same plant but from a different distance.

6.2 Experimental setup and results UCLA8 and UCLA9

The standard test setup for UCLA8 and UCLA9 is to report the average accuracy over 20 random partitions, with 50 % data used for training and 50 % for testing (randomly bisecting each class) [22]. We use the same setup here, except that we report results as an average over 1000 trials to get more reliable statistics.

	Alpha		**Beta**		**Gamma**		
	SVM	NN	SVM	NN	SVM	NN	
Transfer DL [24]	**100.0**	**100.0**	**100.0**	**99.4**	*98.1*	**98.1**	*colour*
Ensemble SVMs [23]	-	-	-	-	**99.5**	-	*colour*
Our approach	98.3	96.7	93.2	92.6	94.3	89.4	*greyscale*
MBSIF-TOP [17]	-	90.0	-	90.7	-	91.3	*greyscale*
AFS-TOP [16]	98.3	91.7	90.1	86.4	94.3	89.4	*greyscale*
LBP-TOP, from [24]	98.3	96.7	88.9	85.8	94.2	84.9	*greyscale*
ELM [10]	-	-	93.8*	-	88.3*	-	*greyscale*
2D+T curvelet [27]	-	88.0†	-	70.0†	-	68.0†	*greyscale*
OTD [18]	87.8*	86.6†	76.7*	69.0†	74.8*	64.2†	*greyscale*
DFS [20]	85.2*	-	76.9*	-	74.8*	-	*greyscale*

Table 2. Comparison to state-of-the-art for the DynTex benchmarks. Superscripts: * indicates a different train-test partitioning and † the use of a nearest centroid classifier.

It can be seen from Table 1 that our proposed approach ranks higher for these two conceptual reorganisations of the database. For UCLA9, we achieve the best result of 98.6 % using an SVM classifier and the second best using a NN classifier, only surpassed by MBSIF-TOP that achieves 98.8 % vs. our 98.3 %. For UCLA8 we achieve 97.8 % compared to the best result by OTD 99.5 % using an SVM classifier, and the second best result of 97.5 % vs 97.8 % using a NN classifier. It can in general be noted that no single approach achieves superior performance on all three UCLA benchmarks.

A confusion matrix for UCLA9 is shown in Figure 3, and we noted that the main cause of error for both UCLA8 and UCLA9 is confusing fire and smoke. There is indeed a similarity in dynamics between these textures in the presence of temporal intensity changes not mediated by spatial movements. Confusions between flowers and plants, as well as between fountain and waterfall, are most likely caused by similarities in the spatial appearance and the motion patterns of these dynamic texture classes.

6.3 Experimental setup and results DynTex

For the DynTex benchmarks, the experimental setup used is leave-one-out cross-validation as in [16, 17, 23, 24]. For this larger and more diverse database, we achieve highly satisfactory results (Table 2). Compared to other methods utilising only grey level information and the same leave-one-out experimental setup, such as AFS-TOP, LBD-TOP and MBSIF-TOP, we achieve the same or better performance on all three benchmarks except for one: when using a NN classifier on the Gamma subset MBSIF-TOP has an accuracy of 91.3 % compared to our 89.4 %. However, we achieve better results for the Beta subset (92.6 % vs. 90.7 %) and substantially better for the Alpha subset (96.7 % vs. 90.0 %). Compared to AFS-TOP we achieve the same results for the Alpha and Gamma subsets using an SVM classifier. There is however a notable improvement for the

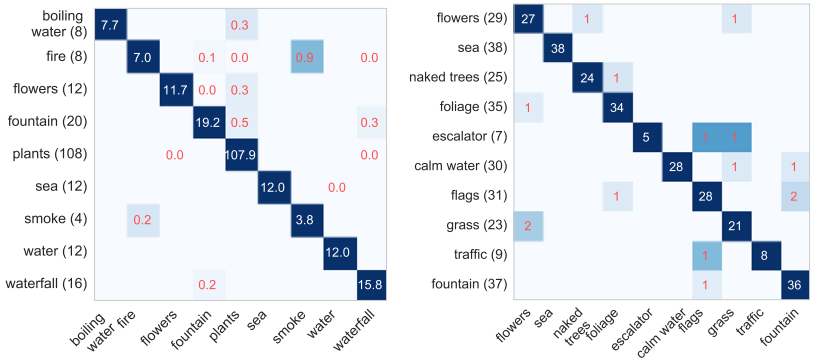

Fig. 3. Left: Confusion matrix UCLA9 averaged over trials (SVM classifier). Right: Confusion matrix DynTex Gamma leave-one-out setting (SVM classifier).

Alpha subset using a NN classifier (96.7 % vs. 91.7 %) and for the Beta subset using both classifiers (SVM: 93.2 % vs. 90.1 %, NN: 92.6 % vs. 86.4 %).

Compared to the original LBP-TOP descriptor, which could be considered a more fair benchmark since we are benchmarking an early version of our approach, we achieve the same performance for the Alpha subset and notably better performance on the Beta (92.6 % error vs. 85.8 %) and Gamma (89.4 % error vs. 84.9 %) subsets using a NN classifier as well as smaller improvements when using an SVM classifier. Interestingly, this is for methods similar to ours in the sense that they collect statistics of local space-time structures and utilise joint binary histogram descriptors. This can be considered a strong validation that the scale-space filters used here capture useful information.

We also show notably better results (in the order of 10-20 p.p.) than those reported from using DFS, OTD and the 2D+T curvelet transform, but since those use a nearest centroid classifier and a different SVM train-test partition, this makes a direct comparison unsure. The best results reported for the three DynTex benchmarks are from transferring deep image features [24] and the ensemble SVM method of Yang et al. [23]. Note that both these approaches utilise colour information which can be highly discriminative for dynamic textures (our approach is straight-forward to extend to colour, which we plan for future work). The latter also combines several descriptors (LBP, SCOP, colour and LDS) which cannot be directly compared to evaluating the performance of a single descriptor.

When inspecting the confusions between different classes for the DynTex benchmarks (Figure 3), the pattern is not very clear, perhaps since this database contains larger intraclass variabilities. We note the largest ratio of misclassified samples for the escalator and traffic classes, which are also the classes with the fewest samples.

7 Summary and discussion

We have presented a new family of video descriptors based on joint histograms of spatio-temporal receptive field responses and evaluated one member of this family on the problem of dynamic texture recognition. This is the first evaluation of using the family of *time-causal scale-space filters* derived in [2] as primitives for video analysis as well as, to our knowledge, the first video descriptor that utilises *joint statistics* of a set of "ideal" (in the sense of derived on the basis of pure mathematical reasoning) spatio-temporal scale-space filters.

Our experimental evaluation on several benchmarks from two widely used dynamic texture databases shows competitive results on the UCLA database and highly competitive results on the larger and more complex DynTex database. For the DynTex benchmarks, we interestingly show improved performance compared to methods similarly modelling statistics of local space-time structure such as local binary pattern based methods [14, 15] and MBSIF-TOP [17], where the latter in contrast to our method utilises filters learned from data. This although temporal causality implies additional constraints on the feature extraction compared to allowing simultaneous access to all video frames. We consider this a strong validation that these spatio-temporal receptive fields are highly descriptive for modelling the local space-time structure, and as evidence in favour of their general applicability as primitives for video analysis.

It should be noted that the method presented here could also be implemented using a non-causal Gaussian spatio-temporal scale-space kernel, which could possibly give somewhat improved results, since at each point in time additional information from the future could also be used. However, a time delayed Gaussian kernel would imply longer temporal delays, which makes it less suited for time critical applications, as well as more computations and larger temporal buffers.

In future work [28], we will generalise the descriptor to colour by considering spatio-chromo-temporal receptive fields and complement the evaluation performed here with an investigation into which receptive field groups works best for dynamic texture recognition as well as how the number of principal components and the number of histogram bins affect the performance. We also plan to broaden the investigation to other video analysis tasks. The theoretical properties of these scale-space filters imply that they could be used to create methods provably invariant or robust to different types of natural image transformations. We see the possibility of integrating time-causal spatio-temporal receptive fields into current video analysis methods as well as using them as primitives for learning higher level features from data.

References

1. Hubel, D.H., Wiesel, T.N.: Brain and Visual Perception: The Story of a 25-Year Collaboration. Oxford University Press (2005)
2. Lindeberg, T.: Time-causal and time-recursive spatio-temporal receptive fields. Journal of Mathematical Imaging and Vision **55** (2016) 50–88

3. Lindeberg, T.: Generalized Gaussian scale-space axiomatics comprising linear scale-space, affine scale-space and spatio-temporal scale-space. Journal of Mathematical Imaging and Vision **40** (2011) 36–81
4. Lindeberg, T.: A computational theory of visual receptive fields. Biological Cybernetics **107** (2013) 589–635
5. Chetverikov, D., Péteri, R.: A brief survey of dynamic texture description and recognition. In: Computer Recognition Systems. Springer (2005) 17–26
6. Linde, O., Lindeberg, T.: Composed complex-cue histograms: An investigation of the information content in receptive field based image descriptors for object recognition. Computer Vision and Image Understanding **116** (2012) 538–560
7. Nelson, R.C., Polana, R.: Qualitative recognition of motion using temporal texture. CVGIP: Image Understanding **56** (1992) 78–89
8. Soatto, S., Doretto, G., Wu, Y.N.: Dynamic textures. In: IEEE International Conference on Computer Vision. Volume 2. (2001) 439–446
9. Ravichandran, A., Chaudhry, R., Vidal, R.: View-invariant dynamic texture recognition using a bag of dynamical systems. In: Computer Vision and Pattern Recognition. (2009) 1651–1657
10. Wang, L., Liu, H., Sun, F.: Dynamic texture video classification using extreme learning machine. Neurocomputing **174** (2016) 278–285
11. Wildes, R.P., Bergen, J.R.: Qualitative spatiotemporal analysis using an oriented energy representation. In: European Conference on Computer Vision. Volume 1843 of LNCS., Springer (2000) 768–784
12. Derpanis, K.G., Wildes, R.P.: Spacetime texture representation and recognition based on a spatiotemporal orientation analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **34** (2012) 1193–1205
13. Gonçalves, W.N., Machado, B.B., Bruno, O.M.: Spatiotemporal Gabor filters: A new method for dynamic texture recognition. arXiv preprint:1201.3612 (2012)
14. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence **29** (2007) 915–928
15. Ren, J., Jiang, X., Yuan, J.: Dynamic texture recognition using enhanced LBP features. In: IEEE International Conference on Acoustics, Speech and Signal Processing. (2013) 2400–2404
16. Hong, S., Ryu, J., Yang, H.S.: Not all frames are equal: Aggregating salient features for dynamic texture classification. Multidimensional Systems and Signal Processing (2016) 1–20
17. Arashloo, S.R., Kittler, J.: Dynamic texture recognition using multiscale binarized statistical image features. IEEE Transactions on Multimedia **16** (2014) 2099–2109
18. Quan, Y., Huang, Y., Ji, H.: Dynamic texture recognition via orthogonal tensor dictionary learning. In: IEEE Int. Conference on Computer Vision. (2015) 7381
19. Schiele, B., Crowley, J.: Recognition without correspondence using multidimensional receptive field histograms. International Journal of Computer Vision **36** (2000) 31–50
20. Xu, Y., Quan, Y., Zhang, Z., Ling, H., Ji, H.: Classifying dynamic textures via spatiotemporal fractal analysis. Pattern Recognition **48** (2015) 3239–3248
21. Ji, H., Yang, X., Ling, H., Xu, Y.: Wavelet domain multifractal analysis for static and dynamic texture classification. IEEE Transactions on Image Processing **22** (2013) 286–299
22. Ghanem, B., Ahuja, N.: Maximum margin distance learning for dynamic texture recognition. In: European Conference on Computer Vision. Volume 6312 of Springer LNCS. (2010) 223–236

23. Yang, F., Xia, G.S., Liu, G., Zhang, L., Huang, X.: Dynamic texture recognition by aggregating spatial and temporal features via ensemble SVMs. Neurocomputing **173** (2016) 1310–1321
24. Qi, X., Li, C., Guoying, Z., Hong, X., Pietikäinen, M.: Dynamic texture and scene classification by transferring deep image features. Neurocomputing **171** (2016) 1230–1241
25. Koenderink, J.J.: Scale-time. Biological Cybernetics **58** (1988) 159–162
26. Péteri, R., Fazekas, S., Huiskes, M.J.: Dyntex: A comprehensive database of dynamic textures. Pattern Recognition Letters **31** (2010) 1627–1632
27. Dubois, S., Péteri, R., Ménard, M.: Characterization and recognition of dynamic textures based on the 2D+T curvelet transform. Signal, Image and Video Processing **9** (2015) 819–830
28. Jansson, Y., Lindeberg, T.: Dynamic texture recognition using time-causal and time-recursive spatio-temporal receptive fields. In preparation (2017)