Stacked Sequential Scale-Space Taylor Context

Carlo Gatta, Francesco Ciompi

Abstract—We analyze sequential image labeling methods that sample the posterior label field in order to gather contextual information. We propose an effective method that extracts local Taylor coefficients from the posterior at different scales. Results show that our proposal outperforms state-of-the-art methods on MSRC-21, CAMVID, eTRIMS8 and KAIST2 datasets.

Index Terms—Contextual modeling, Semantic image labeling, Stacked Sequential Learning

1 Introduction

Semantic image segmentation is the process to simultaneously segment and classify objects in an image. In the last decade contextual information has been used to improve the performance of methods that are based purely on appearance. In this field, the two main approaches are graphical models and sequential methods.

Sequential models are based on the idea of Stacking [1], which provides a sound framework to join appearance and contextual information in a simple manner. The basic idea is that information of object appearance can be complemented by extracting contextual information from the posterior label field. Posteriors are computed by applying a classifier to the appearance features. The contextual information is obtained by sampling the posterior label field from a set of pixels (or regions) around the pixel of interest. However, how to define the set of pixels (also called stencil) is still an open issue, and many approaches define it ad-hoc with respect to the specific problem. An improvement in the theoretical aspect of the stencil definition comes from multi-scale approaches: they require the posterior label field to be filtered with a proper Gaussian function prior to sampling, depending on the distance of the stencil pixel from the pixel of interest.

In this paper, firstly we propose an analysis of different ways of extracting contextual information in sequential methods, which allows us to highlight the benefits and limitations of existing approaches. The main paper contribution is a novel way to extract contextual information based on local Taylor expansion of posterior label field embedded in a scale-space representation. We show that this approach has several advantages w.r.t. prior methods; it allows to: (i) capture the contextual information with less features; (ii) modulate the expressive power of the representation by simply varying the Taylor order and (iii) model spatial label relations in a continuous way. We also provide a complete framework in which our proposal can be implemented and tested on well known datasets.

The paper goes as follows: section 2 briefly introduces related works, section 3 provides an analysis of previous sequential methods, section 4 gives a detailed explanation of our method; sections 5 and 6 provide results and discussion of the method on four datasets; conclusions end the paper.

- *C. Gatta is with the Centre de Visió per Computador, Bellaterra, Barcelona, Spain, 08193. E-mail: cgatta@cvc.uab.es*
- *F. Ciompi is with the Dept. Ciències de la Computació, Universitat Autonoma de Barcelona and the Centre de Visió per Computador, Bellaterra, Barcelona, Spain, 08193.*

2 State of the art

Semantic image segmentation and restoration have been successfully performed by modeling data as Conditional Random Fields (CRFs) [2] and Markov Random Fields (MRFs) [3]. Initially, CRFs were used to handle noisy classification at small scale [4]. Progressively, mid-scale information has been included by hierarchical CRFs [5], [6], [7], together with higher order relationships [8] between super-pixels. Recently, the Random Forest [9] ensemble architecture has been also used to achieve semantic pixel-wise classification and segmentation [10], [11], [12]. Three novel approaches substitute explicit modeling of labels interaction with a learning-based approach [13], [14], [15]. Finally, the global coherence of the segmentation has been improved by adding one or more super-nodes that govern the coexistence of different labels in the same segmentation result [16], [17].

A different set of methods faces the semantic segmentation problem by generating a large set of figure-background segmentations [18]. The combination and/or composition of these segmentations is done by means of different strategies, as minimizing regions overlap while maximizing image coverage, and promoting global consistency of classes.

The idea of sequentially exploiting contextual information in image semantic segmentation has been largely used so far by the Auto-context algorithm [19], [20], the Multi-Scale Stacked Sequential Learning [21] (MS-SSL), the Multi-Scale Context Model [22] (MS-CM), and the Iterative Context Forests [23] (ICF). The idea is based on a seminal paper on the concept of Stacked Sequential Learning [1]. In these approaches, a first classifier generates a posterior class probability using solely appearance features; subsequent classifiers incorporate contextual information by sampling a sparse set of neighbor values [19], [20], or averaging probability from a large set of rectangular regions [23]. To reduce the influence of noise in probability fields due to pure appearance-based prediction, and to extend the sampling to allow efficient long range interaction, some approaches introduce a multi-scale Gaussian filtering (scale-space) prior to sampling [22], [21]. These latter algorithms prove that the multi-scale Gaussian decomposition of posterior label field improves over pure sampling [22] and provides a simple way to implement long range interaction efficiently [21].

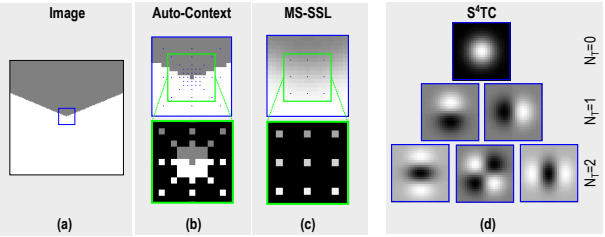

Fig. 1. Different sampling patterns. See text for details.

TABLE 1
# of contextual features as a function of N_c and $D_{\max}$.

Algorithm	Number of contextual features
MS-CM [22]	$8N_c(1+\log_2 D_{\max})$
MS-SSL [21]	$9N_c(1+\log_2 D_{\max})$
Auto-context [19]	$N_c(1+4D_{\max})$
ICF [24]	$\sim N_c D^4_{\max}/8$

3 Analysis of Sequential Methods

The analysis of sequential methods can be structured around two main issues: (1) how the contextual information is gathered from the posterior label field; (2) which kind of relationship between contextual features can be modeled. Figure 1 shows a toy example of posterior label field for a binary problem (a). (b) and (c) show the sampling pattern of Auto-context and MS-SSL (using $\sigma = 4$px), respectively. Sampling pattern of MS-CM is similar to MS-SSL but omitting the central sampling point.

Auto-context. The typical form of sampling is given by the Auto-context algorithm [19] where the density of points decreases with the distance from the center (Figure 1(b)). At each spatial position the posterior of a class can be sampled, or averaged around its 8-neighborhood. Within this sampling scheme, being N_c the number of classes and $D_{\max}$ the maximum distance in pixels from the central one, the number of contextual features grows as $N_c(1+4D_{\max})$. To deal with this number of features, authors perform an implicit feature selection by means of boosting.

MS-SSL / MS-CM. The MS-SSL [21] and the MS-CM [22] have a sampling schema based on a local 8-neighborhood pattern repeated such that the distance of sampling positions to the central pixel grows exponentially, while the posterior is filtered with a Gaussian, having the pixel squared distance as variance. Since both methods use a multi-scale approach in octaves, the number of contextual features grows as $9N_c(1+\log_2 D_{\max})$ and $8N_c(1+\log_2 D_{\max})$ for MS-SSL and MS-CM respectively.

ICF. The ICF method [24] randomly generates a large number of candidates regions where the posterior is averaged to extract contextual information. ICF is basically a sequential version of TextonBoost [4], where the sampling strategy is further extended by extracting contextual features as a function (e.g. difference) of the posterior averaged on rectangular areas. Being N the number of image pixels, even using the simplest "rectangular-based" sampling, the number of potential rectangles is $N_c N^2/2$, which is clearly intractable per se. To deal with such a high number of features, the classical strategy is to perform an implicit feature selection by boosting or use random forest as a basic classifier.

Table 1 summarizes the number of contextual features for all the analyzed methods. To properly compare [24] to other methods, we define the maximal distance $D_{\max}$ as the diagonal of the image that, assuming square images, is $D_{\max} = \sqrt{2N}$. This results in $N_c D^4_{\max}/8$ features. It is clear that, since all the methods sample the posterior label field, the number of contextual features grows linearly with N_c. However, to perform long range interaction up to a certain distance $D_{\max}$, different methods require different number of contextual features. The methods requiring less contextual features are MS-CM and MS-SSL, while the one that requires more is ICF. Thanks to the octave-based multi-scale decomposition of MS-SSL and MS-CM, the number of features grows logarithmically with the desired maximal distance of interaction. On the contrary, pure sampling methods require, for the same level of long-range interaction, more contextual features. Moreover, the lack of Gaussian (or any form of) filtering prior to sampling, induces a loss of information and makes the algorithm more sensitive to noisy posterior label fields.

Regarding label interaction, differently from graphical models, sequential methods do not impose an explicit modeling; all of the above mentioned sequential methods use contextual information as pure *features*, allowing to delegate the classifier to learn their relationship. In our humble opinion, this is the most important advantage of sequential models over graphical models. However, within sequential approaches, we think that mere sampling of the posterior label field poorly models the context. Our hypothesis is that the extraction of scale-space shape descriptors of the posterior label field allows better performance while requiring less contextual features. We implement this strategy by means of the Scale-Space Taylor coefficients.

4 Stacked Sequential Scale-Space Taylor Context (S⁴TC)

Our method is founded on the Scale-Space Taylor Context descriptor embedded in the Stacked Sequential Learning architecture [1].

4.1 Scale-Space Taylor Context

A data observation at position $\mathbf{q}$ can be described by its feature vector $\mathbf{x}(\mathbf{q}) \in \mathbb{R}^{N_\mathbf{x}}$ and its label $c \in Y$. In a classification problem with N_c classes, let us assume the existence of a functional $f : \mathbb{R}^{N_\mathbf{x}} \rightarrow [0,1]^{1\times N_c}$ that provides the (pseudo-) probability for a feature vector to belong to each one of the classes. In correspondence to a position $\mathbf{q}$ of the data domain Θ, a vector $\mathbf{p}(\mathbf{q}) =$

$f(\mathbf{x}(\mathbf{q})) \in [0,1]^{1\times N_c}$ is obtained, such that the assigned label can be computed as $\hat{c}(\mathbf{q}) = \arg_c \max(p_c(\mathbf{q}))$. In the rest of the paper we refer to the set of vectors $\mathbf{p}$ computed over the data domain Θ as $\mathcal{P} \in [0,1]^{\Theta\times N_c}$, which represents the spatial distribution of the probabilities, for each class, over the domain. We refer to $\mathcal{P}$ as the *posterior label field*.

The value of $\mathcal{P}$ in a region Ω centered in the position $\mathbf{q}_0$ can be approximated, following Lindeberg [25], by means of the Taylor expansion, whose coefficients $T^{\sigma_T}_{v,z}$ for images, omitting constants, are:

$$T^{\sigma_T}_{v,z}(\mathbf{q}_o) = \left.\frac{\partial^{v+z}\mathcal{P} * G(0,\sigma_T)}{\partial^v x \partial^z y}\right|_{\mathbf{q}_o}, \quad (1)$$

where $N_T = v + z$ is the expansion order, with $v \geq 0$ and $z \geq 0$, and $G(0,\sigma_T)$ indicates a Gaussian filter of zero mean value and standard deviation σ_T. Given the desired maximum expansion order $N_T^{\max}$, we compute the coefficients $T^{\sigma_T}_{v,z}$ for each order $N_T \leq N_T^{\max}$, each scale σ_T and for each class c, which are concatenated to form the contextual feature vector $\mathbf{x}_T$. Figure 1 (d) shows the Taylor "sampling pattern" in form of spatial filters, up to $N_T^{\max} = 2$. This representation is also known as N-jet [26]; however, it has not to be confused to the N-jet applied directly on image data. Here we are modeling the context as the spatial relationship of values in the posterior label field $\mathcal{P}$.

With the proposed approach, given N_c classes and N_σ scales, the Taylor-based feature vector $\mathbf{x}_T$ is obtained by concatenating N_c times N_σ vectors, thus the length of $\mathbf{x}_T$ grows as

$$\frac{(N_T^{\max}+2)(N_T^{\max}+1)}{2} N_c N_\sigma. \quad (2)$$

This approach has the advantage to allow modeling the neighborhood complexity by simply increasing the maximum Taylor order $N_T^{\max}$.

In all the experiments we define the Scale-Space by means of a set of Gaussian standard deviations $\sigma_T = 2^\iota$, where $\iota \in \{0,1,\ldots,N_\sigma - 1\}$, i.e. in octaves, and N_σ is the number of scales. If we desire to have long range interactions up to a certain distance $D_{\max}$, the length of our contextual feature grows as

$$\frac{(N_T^{\max}+2)(N_T^{\max}+1)}{2} N_c (1 + \log_2 D_{\max}). \quad (3)$$

Comparing equation (3) to Table 1, it can be noted that our approach has less contextual features than all the other ones by setting $N_T^{\max} \leq 2$.

The maximum order of the expansion $N_T^{\max}$ can be selected according to the properties of the problem. This is due to the different meaning that each order assumes. The 0th order term is merely the sampling of a Gaussian blurred version of $\mathcal{P}$ in the central pixel position, thus providing information on the local probability of the classes. The 1st order terms provide information about transitions of the values, i.e. the presence of edges on the posterior label field. The directionality of the transition is completely encoded by the two terms in a continuos way. Using the 1st order terms, the proposed method can deal with spatial relationships between classes, coding their relative position. The 2nd order terms can compactly codify shapes like circular blobs, ellipses of any orientation and eccentricity. This allows to codify multiple transitions between classes in an area. The 3rd order terms (and higher) codify far more complex shapes and their utility in modeling semantic spatial relationship should be estimated case by case.

Algorithm 1 Training

Require: $\mathbf{D}_{train}$, Y_{train}, N_i, γ_A, γ_S, γ_T
Ensure: $\{f_i\}$ {set of trained functionals for each i}
1: $\mathbf{X}_A = \gamma_A(\mathbf{D}_{train})$
2: $\mathbf{x}_T^{(1)} = \emptyset$;
3: **for** i = 1:N_i **do**
4: $\quad\mathbf{x}_E^{(i)} = [\mathbf{X}_A\ \mathbf{x}_T^{(i)}]$
5: $\quad\tilde{\mathbf{x}}_E^{(i)} = \gamma_S(\mathbf{x}_E^{(i)})$
6: $\quad$TRAIN $f_i|_{\{\tilde{\mathbf{x}}_E^{(i)}, \gamma_s(Y_{train}^{(i)})\}}$
7: $\quad$**if** $i < N_i$ **then**
8: $\quad\quad\mathcal{P}^{(i)} = f_i(\mathbf{x}_E^{(i)})$
9: $\quad\quad\mathbf{x}_T^{(i+1)} = \gamma_T(\mathcal{P}^{(i)})$
10: $\quad$**end if**
11: **end for**

4.2 Training and Inference in SSL

To the sake of completeness we present training and testing algorithms of our method, which are based on an iterated version of SSL.

Training. The procedure for training S^{4}TC is described in Algorithm 1.

The requirements are: a dataset of training data $\mathbf{D}_{train}$, together with their corresponding labels Y_{train}; a functional γ_A for extracting appearance based features; a functional γ_T for extracting the scale-space Taylor coefficients, as in equation (1); a functional γ_S for randomly subsampling a set of data. Following the procedure of the SSL strategy [1], the subsampling operator γ_S must ensure that at different iterations the classifier is trained on disjoint subsets, i.e. $\gamma_S(\mathbf{x}_E^{(i)}) \cap \gamma_S(\mathbf{x}_E^{(j)}) = \emptyset$ for each pair of iterations $i \neq j$.

The training procedure consists in an iterative process where a functional f_i is trained at each iteration $i = (1,\ldots,N_i)$, fed by the concatenation of data (or appearance) and Taylor contextual features $\mathbf{x}_T$, namely by the extended set $\mathbf{x}_E$ (step 4). Before each training step, a subset of $\mathbf{x}_E$ is randomly selected as $\tilde{\mathbf{x}}_E = \gamma_S(\mathbf{x}_E)$ (step 5). The set $\tilde{\mathbf{x}}_E$, together with the corresponding labels $\gamma_s(Y_{train}^{(i)})$, is then used to train a functional f_i (step 6), successively applied to the whole training dataset $\mathbf{x}_E$ to compute the posterior label field $\mathcal{P}$. Finally, the contextual features $\mathbf{x}_T$ are extracted by means of the function γ_T (step 10), to be used in the next iteration. It is worth noting that contextual information is not used to train the functional f_1 (step 2). The output of the training procedure is the set of functionals $\{f_i\}$.

Algorithm 2 Inference

Require: $\{f_i\}$, D_{test}, γ_A, γ_T
Ensure: $\mathcal{Y}$ {label field for image D_{test}}
1: $\mathbf{x}_A = \gamma_A(D_{test})$
2: $\mathbf{x}_T^{(1)} = \emptyset$;
3: **for** i = i:N_i **do**
4: $\quad\mathbf{x}_E^{(i)} = [\mathbf{x}_A\ \mathbf{x}_T^{(i)}]$
5: $\quad\mathcal{P}^{(i)} = f_i(\mathbf{x}_E^{(i)})$
6: $\quad$**if** $i < N_i$ **then**
7: $\quad\quad\mathbf{x}_T^{(i+1)} = \gamma_T(\mathcal{P}^{(i)})$
8: $\quad$**end if**
9: **end for**
10: $\mathcal{Y} = \arg_c \max \mathcal{P}_c^{(N_i)}$

Inference. The procedure for inference in S^{4}TC is described in Algorithm 2. Given a test datum D_{test} and the set of trained functionals $\{f_i\}$, the appearance based features are first extracted (step 1) and joined with contextual features $\mathbf{x}_T$ (step 4); as for training, at the first iteration no contextual features exist (step 2). Applying f_i to $\mathbf{x}_E^{(i)}$ at each iteration produces the posterior $\mathcal{P}^{(i)}$ (step 5), which is used to construct Taylor contextual features $\mathbf{x}_T$, for the next iteration (step 7). The output label field is computed as the maximum a posteriori over $\mathcal{P}^{(N_i)}$ (step 11).

5 Experiments

In this section we describe the experiments and datasets used to evaluate the performance of S^{4}TC. We selected four datasets: the MSRC-21 is largely used to evaluate context-aware methods; the eTRIMS8 and CAMVID are datasets presenting highly structured label distributions; and finally, the KAIST2 dataset is used to perform a difficult in-painting task.

5.1 MSRC-21

In this section we provide the results of our method on the MSRC-21 dataset [4].

Appearance features. We use features composed by L*a*b* color, smoothed with a Gaussian filter at several scales $\sigma_A = \{.25, .5, 1, 2, 4, 8\}$ px; color SIFT on L*a*b* channels at two patch size (24×24 and 48×48 px); Gaussian weighted ($\sigma_G = 8$ px) color histogram of the a*b* channels (with 169 bins); we also use a position prior learned from the ground truth training labels, estimated with a Kernel Density Estimator ($\sigma_K = 4$ px).

Computing $\mathcal{P}$. In order to compute $\mathcal{P}$ we adopt the Error-Correcting Output Codes framework [27] with *One Against All* technique, which allows to decompose the multi-class classification problem into a set N_c of binary problems. A matrix of codes $\mathbf{M} \in \{-1, 1\}^{N_c \times N_c}$ is obtained, and the value of the pseudo-probability is $p(c) \propto \exp^{-\alpha d_{ECOC}(c)}$, where $d_{ECOC}(c)$ is the Euclidean distance between the code obtained by concatenating the output of all the binary classifiers and the ECOC code corresponding to the label class c, using $\alpha = \ln(N_c)/\sqrt{N_c}$. For this dataset, the binary classifier is Support Vector Machine (SVM) with intersection kernel [28]. In order to train SVMs, both data features and contextual features are whitened and rescaled to the range $[0, 1]$ using the sigmoid function. At each iteration, the SVM regularization parameter C has been tuned by cross-folding optimizing the accuracy of the output of the ECOC matrix. To have a tractable tuning process, the value of C during tuning is shared by all the *One Against All* SVM classifiers. Training has been performed with 5000 samples per class using $N_i = 5$.

Table 2 shows the Global algorithm accuracy and the Average class sensitivity from iteration 1 to 5, and for different Taylor orders $N_T^{\max}$. At a first sight, it could be surprising that, even with just the 0th order Taylor coefficient, the algorithm is able to perform well. However, this can be explained by the fact that the scale-space filtering allows the classifier to learn, at least, the most probable class combinations at different scales, somehow learning the "scale-space co-occurence" of labels. Nonetheless, first and second order coefficients, as expected, consistently improve the results. When we add the 3rd order contextual information ($N_T^{\max} = 3$), the performance drops significantly. This can be explained by the following sentence, excerpted from [29]: "[... we showed that SVMs can indeed suffer in high dimensional spaces where *many* features are *irrelevant*.]". The 3rd order Taylor expansion indeed requires many features (672). In order to quantify the relevance of 3rd order features, we used the F-score [30], on the Taylor coefficients produced at the second iteration of the algorithm. Results show that 3rd order Taylor coefficients are much less relevant than lower orders, thus supporting our hypothesis. This limitation of SVM can be overcome by using a feature selection [29] or features weighting algorithm [31]; however, this is out of the scope of the present paper and, considering the results on eTRIMS8 and KAIST2, we do not expect it could provide clear performance improvement.

The biggest improvement is provided by the second iteration, where contextual features are used for the first time by the algorithm. Subsequent iterations improve both performance measures, but the increment at each further iteration is smaller than the previous. This behavior is typical of SSL-based approaches, which exhibit similar performance curves for other datasets [32], [22]. Table 3 shows the behavior of the algorithm while adding scales, using $N_T^{\max} = 2$. It is clear that the long-range interaction is a key component of the method; the performance constantly increases while adding scales. Table 4 shows a comparison between three sequential-based methods plus four state-of-the-art methods and S^{4}TC. The row "Baseline-ST^{4}C" shows the result of the first iteration of our algorithm, where only appearance features are used. The improvement due to our method is significantly large for almost all classes (+22.5% on average), and it is not negligible even for classes that had already a good performance in the baseline result. Our method presents the highest average sensitivity among

TABLE 4
MSRC-21 dataset. Average provides the per-class sensitivity. Global gives the percentage pixel accuracy.

	Global	Average	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat
MS Auto-context [20]	78	68	53	**97**	83	70	71	98	75	64	74	64	88	67	46	32	92	61	89	59	66	64	13
Auto-context [19]	77	69	69	96	87	78	80	95	83	67	84	70	79	47	61	30	80	45	78	68	52	67	27
MS-SSL [21]	83	78.6	63	93	**94**	**92**	89	96	96	69	86	78	92	86	68	56	85	60	82	86	62	78	42
Fully conn. CRF [33]	86	78.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Hierar. w/ CO [16]	**87**	77	**82**	95	88	73	88	**100**	83	**92**	**88**	**87**	88	**96**	**96**	27	85	37	**93**	49	**80**	65	20
Harmony Pot. [17]	83	80	66	87	84	81	83	93	81	82	78	86	**94**	**96**	87	48	90	**81**	82	82	75	70	52
D-sampl. L+G [34]	79	78	50	83	87	81	84	90	**97**	72	75	79	90	95	79	52	**97**	**81**	80	**89**	51	64	**60**
Baseline-ST^{4}C	66	58	36	92	71	59	64	89	81	59	72	46	70	57	41	35	61	29	71	44	42	59	30
ST^{4}C ($N_T^{\max}=2$, $N_i=5$)	84	**80.5**	67	95	**92**	91	**90**	95	96	73	**88**	76	**94**	90	76	**57**	84	69	82	**89**	60	**84**	44

TABLE 2
Global Accuracy/Average sensitivity on MSRC-21 varying $N_T^{\max}$ and N_i.

$N_T^{\max}$	$N_i=1$	$N_i=2$	$N_i=3$	$N_i=4$	$N_i=5$
0	66.2/57.6	78.3/73.3	80.4/76.6	81.6/78.0	81.9/78.8
1	66.2/57.6	79.9/75.0	82.0/78.0	83.1/79.2	83.6/80.2
2	66.2/57.6	80.9/76.2	82.7/78.8	83.7/79.7	84.0/80.5
3	66.2/57.6	79.9/74.8	81.7/76.8	82.2/77.3	82.6/77.9

TABLE 3
Global accuracy and average sensitivity using $N_T^{\max}=2$, $N_i=5$, while varying the number of scales on MSRC-21.

$\sigma_T^{\max}$	1	2	4	8	16	32	64	128
Global	67.4	68.5	71.1	74.6	77.3	79.6	82.3	84.0
Average	60.2	61.9	65.1	69.0	72.5	75.4	78.8	80.5

all the compared algorithms, and the best sensitivity on 7 classes, with a remarkable 57% for the "bird" class.

Figure 2 shows 5 results of our method.

In the first row we show the test images, in the second row the corresponding classification output at iteration $i=1$ (when no contextual information is used); in the third row we show the classification output at iteration $i=5$; in the fourth row we also depict the manual ground truth labeling, while in the last row we show the pixel-wise entropy map computed over $\mathcal{P}^{(5)}$ at the last iteration. As a first observation, it is interesting to note the effectiveness of the proposed architecture that, thanks to the use of contextual features, dramatically improves the labeling from iteration 1 to 5; in particular, even in cases with completely unstructured initial results (a, b, c), S^{4}TC is able to achieve an accurate final labeling. Furthermore, in cases where the final labels are less accurate (e), the method still provides a semantically consistent solution, with configurations of regions in the image seen during the training procedure. The stability (or instability) of such configurations is shown by the entropy map, which in most of the cases clearly outline the main regions of the image with low values, keeping uncertainty around the label transitions.

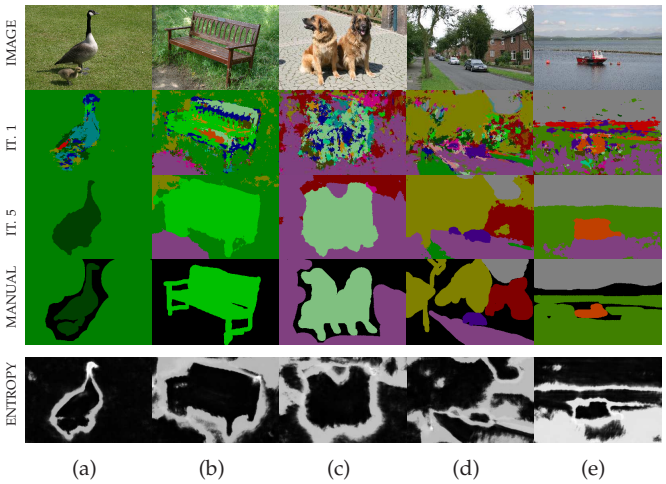

Fig. 2. Classification results for the MSRC-21 dataset.

5.2 eTRIMS8

The eTRIMS8 dataset consists of 60 pictures of buildings; for each image, accurate pixel-wise manual annotation of the following categories is provided[1]: *building, car, door, pavement, road, sky, vegetation, window*. In this dataset, the labels spatial distribution is highly structured, making it a good benchmark to evaluate the proposed method ability to represent context. We mimic the experiment in [35], thus performing a five-folds cross-validation. To get a fair comparison, we tuned our appearance features to obtain a similar baseline result of approximately 66% accuracy as in [35].

Appearance features. We use zero, first and second order derivatives over the three L*a*b* chromatic channels of the image, over several scales with $\sigma_A = \{1, 2, 4, 8, 16, 32\}$ and a dense SIFT descriptor on 16×16 patches over the L*a*b* channels.

Computing $\mathcal{P}$. The base classifier is a SVM with Intersection Kernel. The functionals f_i are obtained by combining 8 classifiers in a *One Against All* fashion within the ECOC framework. Training, tuning and inference is done as for the MSRC-21 dataset.

1. www.ipb.uni-bonn.de/projects/etrims_db/

(a) (b) (c) (d) (e)

Fig. 3. Classification result on eTRIMS8 from iteration 1 (a) to iteration 4 (d) and ground truth labels (e).

Figure 3 shows an example of the results on a test image.

As it can be noticed, the appearance based result (a) is noisy and presents incorrectly classified regions. The second iteration, which is the first accessing the contextual features, corrects most of erroneous labels that do not match the learned contextual configuration (b), e.g. the wrongly classified *door* pixels at the second floor of the house. Subsequent iterations refine the result providing a sharp pixel-wise classification. It is worth to note how the system learns the expected relative spatial position of the classes: the pavement is correctly classified only by its context, i.e. its spatial position w.r.t. the road, vegetation and the building.

Table 5 shows the accuracy for different Taylor orders $N_T^{\max}$. Increasing the Taylor order up to $N_T^{\max} = 3$ improves the algorithm performance, reaching up to 83.4%. However, with $N_T^{\max} = 3$ there is no increment in accuracy with respect to $N_T^{\max} = 2$, showing that the additional information provided by the third order derivative is not useful to model the contextual spatial relationship between classes. This is not surprising since third order derivatives capture structures which are not (or rarely) present in any label spatial pattern in the eTRIMS8 dataset. Table 5 also shows a numerical comparison with the Hierarchical CRFs method in [7], the Multi-Scale Stacked Sequential Learning (MS-SSL) in [21] and the facade parsing algorithm in [35]. It has to be noted that the algorithm in [35] is specifically designed to perform facade parsing, so that it uses a-priori knowledge on architectural properties of facades. With respect to the MS-SSL [21], which has been learned with the same appearance features and classifier, S^{4}TC provides a higher accuracy with less contextual features.

5.3 CAMVID

In this section we present the results on the CamVid dataset following the train/test split proposed in [36], using 468 images for training and 233 for the test. The settings of our algorithm are exactly the same as the ones used for the MSRC-21 dataset for both appearance and contextual features. As it can be noted in Table 6, our method provides the best Average sensitivity w.r.t. to all previous methods. It has to be noted that while we use solely appearance and context information, some of the state-of-the-art methods use additional structure-from-motion and/or depth maps information.

TABLE 5
eTRIMS8 performance comparison.

	Building	Car	Door	Pavement	Road	Sky	Vegetation	Window	Baseline	Pixel acc.
Hier. CRF [7]	67	36	14	**85**	53	80	78	**80**	65.8	69.0
MS-SSL [21]	82	57	68	55	68	99	92	79	**66.4**	81.9
Facade parsing [35]	**90**	66	20	35	47	90	83	75	65.8	81.9
RF approach [24]	71	**77**	71	70	73	95	90	64	-	75.1
S^{4}C, $N_T^{\max} = 0$	76	67	73	49	65	99	92	78	**66.4**	79.4
S^{4}C, $N_T^{\max} = 1$	79	55	69	50	70	99	92	78	**66.4**	80.8
S^{4}C, $N_T^{\max} = 2$	83	54	**75**	57	80	**99**	**92**	78	**66.4**	**83.4**
S^{4}C, $N_T^{\max} = 3$	84	51	73	55	81	99	92	78	**66.4**	**83.4**

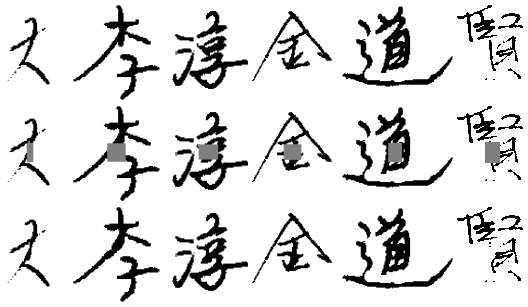

Fig. 4. Some exemplar results of S^{4}TC on KAIST2.

5.4 KAIST2 dataset

In this section we present the results of the proposed method on the problem of in-painting chinese characters using the KAIST2 dataset. We reproduced exactly the same experiment proposed in [15], using 300 binary images for training and 100 images for testing. The experiment is meant to measure the ability of a method to learn the implicit structure of chinese characters in order to recover the original from the corrupted one. Figure 4 (first row) shows some examples of chinese characters from the dataset, together with its corrupted version (second row), where the gray square represents the "deleted" area; last row shows the reconstruction by S^{4}TC.

The experiment is extremely interesting since there is no way to use local appearance features, due to the fact that the region to be in-painted is totally erased. Moreover, the chinese character structure is complex, so that the problem is not of mere interpolation, but it is more related with hallucinating the correct (or most probable) structure likely to match the visible part of the character. In this experiment we iterate the algorithm up to $N_i = 4$ iterations.

Computing $\mathcal{P}$. For this experiment, with the aim to be directly comparable to [15], we use Random Forest with 100 trees, using 3000 samples per class at each iteration. Since it is a binary classification problem, the posterior provided by Random Forest is used to build $\mathcal{P}$ without the need of the ECOC framework.

Table 7 shows the accuracy of S^{4}TC varying the Taylor

TABLE 6
Performance comparison for the CAMVID dataset.

	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalks	Bicyclist	Average	Global
Sturgess et al. [37]	84.5	72.6	**97.5**	72.2	34.1	95.4	34.2	45.7	8.1	77.6	28.5	59.2	83.8
Zhang et al. [38]	**85.3**	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	22.0	38.1	28.7	55.4	82.1
Floros et al. [39]	80.4	76.1	96.1	86.7	20.4	95.1	47.1	47.3	8.3	79.1	19.5	59.6	83.2
Ladicky et al. [40]	81.5	**76.6**	96.2	**78.7**	40.2	93.9	43.0	47.6	14.3	**81.5**	33.9	62.5	83.8
Tighe et al. [41]	83.1	73.5	94.6	78.1	**48.0**	96.0	58.6	32.8	5.3	71.2	**45.9**	62.5	**83.9**
Baseline-ST^{4}C	47.3	57.6	93.3	67.8	35.5	85.8	69.8	39.7	39.2	69.7	24.0	57.2	70.5
ST^{4}C ($N_T^{\max}=0$, $N_i=4$)	55.6	72.8	93.6	79.2	36.9	89.2	74.2	54.2	40.6	76	42.4	65.0	76.6
ST^{4}C ($N_T^{\max}=1$, $N_i=4$)	54.3	71.9	93.4	79	43.4	90.9	72.5	51.8	43.3	72.6	39.3	64.7	76.2
ST^{4}C ($N_T^{\max}=2$, $N_i=4$)	55.7	71.9	93.6	75.5	43.2	91.4	**74.2**	**53.5**	**47.4**	74.2	41.8	**65.7**	76.9
ST^{4}C ($N_T^{\max}=3$, $N_i=4$)	53.8	71.1	93.5	75.5	40.6	91.7	74.8	53.6	48.6	74.3	39.6	65.2	76.4

TABLE 7
Accuracy on KAIST2 varying $N_T^{\max}$ and N_i.

	$N_i=1$	$N_i=2$	$N_i=3$	$N_i=4$
$N_T^{\max}=0$	57.07	73.90	73.99	73.74
$N_T^{\max}=1$	57.07	74.68	74.05	74.45
$N_T^{\max}=2$	57.07	77.92	79.33	79.72
$N_T^{\max}=3$	57.07	77.57	77.96	78.37

TABLE 8
KAIST2 performance comparison.

	DTF [15]	MS-SSL [21]	SLP [14]	S^{4}TC
Accuracy	76.01	76.9	78.08	**79.72**
# Params	64	54	64 (16×4)	24

order at different iterations.

Using $N_T^{\max}=0$ the contextual information is merely a low pass filtering of neighbor character strokes; in this case the accuracy is 73.74%. When information on first derivatives is added, the performance increases to 74.45%, but when second order information is included, the system accuracy reaches 79.72%. This increment in performance is due to the fact that second order derivatives can encode ellipses and blobs, approximating strokes, dots and holes in chinese characters. It is also worth noticing that using up to the third order derivative does not add useful contextual information.

Table 8 shows comparison between S^{4}TC and three state-of-the-art methods for KAIST2 in terms of accuracy and number of required parameters. As it can be noted, using the S^{4}TC optimal configuration ($N_T^{\max}=2$, $\sigma_T^{\max}=4$) allows to outperform previous results in terms of accuracy, while using much less parameters[2]. This clearly shows that the Scale-Space Taylor coefficients are both expressive and compact.

2. Following equation (3), the number of contextual features for S^{4}TC results in 48. However, since the Taylor context is linear, and the problem is binary, half of the features are linearly correlated, so that can be safely removed.

6 Discussion

The presented results confirm that modeling the shape of posterior label field within a scale-space decomposition provides better results and less contextual features than classical sequential methods, such as the MS-SSL and, in general, better than state-of-the-art graphical methods.

In all experiments, the maximum Taylor degree that provided best results is $N_T^{\max}=2$. This is not surprising, since local average, gradient and Hessian matrix are well known effective descriptors of local properties of an image. The use of a very basic descriptor to encode local shape of the posterior label field demonstrated to be a simple yet efficacious strategy to extract contextual information. Higher Taylor orders seem not to provide relevant information; this can be explained by the fact that, in natural images, there are no complex spatial label patterns that require the use of 3rd order Taylor coefficients. Nonetheless, for other kinds of sequential data, e.g. for audio signals, the 3rd order Taylor coefficients could be useful to model alternating patterns.

The scale-space approach offers a very efficient way of representing the data at different sizes, so that it allows to learn expected "size" of objects in terms of posterior label field instead of their image appearance; and also their co-occurence at different scales.

Using $N_T^{\max}=2$, S^{4}TC requires a number of contextual features that grows as $6N_c(1+\log_2 D_{\max})$. Comparing this to other methods in Table 1, it can be noted that S^{4}TC requires less contextual features that any other sequential algorithm, while outperforming them in all the experiments. The most relevant aspects of having the lowest number of contextual features is that we can train powerful classifiers providing all the contextual information *at once*, allowing to learn complex relationships between labels. We believe that implicit or explicit feature selection could potentially hinder the possibility of learning less frequent (or more complex) label configurations.

As for other sequential methods, it can be demonstrated that S^{4}TC monotonically decreases the training error trough iterations [19]. In S^{4}TC, being based on

the Stacked Sequential scheme, the modeling of interaction between appearance and context is delegated to the classifier. This appears to be a relevant issue, as demonstrated by the increasing interest in learning label interaction also in graphical models [13], [14], [15].

7 Conclusion

In this paper we presented Stacked Sequential Scale-Space Taylor Context. S^{4}TC outperforms state-of-the-art algorithms on four datasets: MSRC-21, eTRIMS8, CAMVID and KAIST2. Due to its modularity, we expect the method to be able to deal with other kind of sequential data, and generalize to higher dimensional data (volumes or videos) due to its good scalability.

Acknoledgments

Authors want to thank the anonymous reviewers, J. van de Weijer and F. Vilariño for valuable comments and discussions. The work of C. Gatta is supported by MICINN under a Ramon y Cajal Fellowship. This paper is dedicated to the memory of my father Annibale Gatta.

References

[1] W. W. Cohen, "Stacked sequential learning," in *International Joint Conference on Artificial Intelligence*, 2005, pp. 671–676.
[2] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," in *ICML*, 2001, pp. 282–289.
[3] D. Geman and S. Geman, "Stochastic relaxation, gibbs distribution, and the bayesian restoration of images," *TPAMI*, vol. 6, pp. 721–741, 1984.
[4] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *Int. J. Comput. Vision*, vol. 81, no. 1, pp. 2–23, Jan. 2009.
[5] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical crfs for object class image segmentation," in *ICCV*. IEEE, 2009, pp. 739–746.
[6] Q.-X. Huang, M. Han, B. Wu, and S. Ioffe, "A hierarchical conditional random field model for labeling and segmenting images of street scenes," in *CVPR*. IEEE, 2011, pp. 1953–1960.
[7] M. Ying Yang and W. Forstner, "A hierarchical conditional random field model for labeling and classifying images of man-made scenes," in *IEEE/ISPRS workshop on Computer Vision for Remote Sensing of the Environment*, 2011, pp. 196–203.
[8] P. Kohli, L. Ladicky, and P. H. S. Torr, "Robust higher order potentials for enforcing label consistency," *IJCV*, vol. 82, no. 3, pp. 302–324, 2009.
[9] L. Breiman, "Random forest," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[10] A. Montillo, J. Shotton, J. M. Winn, J. E. Iglesias, D. N. Metaxas, and A. Criminisi, "Entangled decision forests and their application for semantic segmentation of ct images," in *IPMI*, ser. LNCS, vol. 6801. Springer, 2011, pp. 184–196.
[11] P. Kontschieder, S. Rota Bulò, A. Criminisi, P. Kohli, M. Pelillo, and H. Bischof, "Context-sensitive decision forests for object detection," in *NIPS 25*, 2012, pp. 440–448.
[12] P. Kontschieder, S. Rota Bulò, H. Bischof, and M. Pelillo, "Structured class-labels in random forests for semantic image labelling," in *ICCV*, 2011, pp. 2190–2197.
[13] F. Ciompi, O. Pujol, and P. Radeva, "A meta-learning approach to Conditional Random Fields using Error-Correcting Output Codes," *IEEE ICPR*, pp. 710–713, 2010.
[14] S. Rota Bulò, P. Kontschieder, M. Pelillo, and H. Bischof, "Structured local predictors for image labelling." in *CVPR*. IEEE, 2012, pp. 3530–3537.
[15] S. Nowozin, C. Rother, S. Bagon, T. Sharp, B. Yao, and P. Kholi, "Decision tree fields," in *ICCV*, 2011, pp. 1668–1675.
[16] L. Ladicky, C. Russell, P. Kohli, and P. H. S. Torr, "Graph cut based inference with co-occurrence statistics," in *ECCV (5)*, ser. LNCS, vol. 6315. Springer, 2010, pp. 239–253.
[17] X. Boix, J. M. Gonfaus, J. van de Weijer, A. D. Bagdanov, J. S. Gual, and J. Gonzàlez, "Harmony potentials - fusing global and local scale for semantic image segmentation," *IJCV*, vol. 96, no. 1, pp. 83–102, 2012.
[18] J. a. Carreira, F. Li, and C. Sminchisescu, "Object recognition by sequential figure-ground ranking," *Int. J. Comput. Vision*, vol. 98, no. 3, pp. 243–262, Jul. 2012.
[19] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3d brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1744–1757, 2010.
[20] J. Jiang and Z. Tu, "Efficient scale space auto-context for image segmentation and labeling," in *CVPR*. IEEE, 2009, pp. 1810–1817.
[21] C. Gatta, E. Puertas, and O. Pujol, "Multi-scale stacked sequential learning," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2414–2426, 2011.
[22] M. Seyedhosseini, R. Kumar, E. Jurrus, R. Giuly, M. H. Ellisman, H. Pfister, and T. Tasdizen, "Detection of neuron membranes in electron microscopy images using multi-scale context and radon-like features," in *MICCAI*, ser. LNCS, vol. 6891, 2011, pp. 670–677.
[23] D. Munoz, J. A. D. Bagnell, and M. Hebert, "Stacked hierarchical labeling," in *ECCV*, 2010, pp. 57–70.
[24] B. Fröhlich, E. Rodner, and J. Denzler, "Semantic segmentation with millions of features: Integrating multiple cues in a combined random forest approach," in *ACCV*, 2012, pp. 218–231.
[25] T. Lindeberg, "Scale-space theory: A basic tool for analysing structures at different scales," *Journal of Applied Statistics*, pp. 224–270, 1994.
[26] L. Florack, B. T. H. Romeny, M. Viergever, and J. Koenderink, "The gaussian scale-space paradigm and the multiscale local jet," *International Journal of Compyer Vision*, vol. 18(1), pp. 61–75, 1996.
[27] T. G. Dietterich and G. Bakiri, "Solving Multiclass Learning Problems via Error-Correcting Output Codes," *JAIR*, vol. 2, pp. 263–286, 1995.
[28] S. Maji, A. C. Berg, and J. Malik, "Efficient classification for additive kernel svms," *TPAMI*, vol. 35, no. 1, pp. 66–77, 2013.
[29] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik, "Feature selection for svms," in *Advances in Neural Information Processing Systems 13*. MIT Press, 2000, pp. 668–674.
[30] Y.-W. Chen and C.-J. Lin, "Combining SVMs with Various Feature Selection Strategies," in *Feature Extraction*, ser. Studies in Fuzziness and Soft Computing. Springer Berlin Heidelberg, 2006, vol. 207, ch. 13, pp. 315–324.
[31] S. Zhang, M. M. Hossain, M. R. Hassan, J. Bailey, and K. Ramamohanarao, "Feature weighted svms using receiver operating characteristics," in *Proceedings of the SIAM International Conference on Data Mining, SDM*. SIAM, 2009, pp. 497–508.
[32] Z. Kou and W. W. Cohen, "Stacked graphical models for efficient inference in markov random fields," in *SDM*. SIAM, 2007.
[33] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NIPS*, 2011, pp. 109–117.
[34] A. Lucchi, Y. Li, X. B. Bosch, K. Smith, and P. Fua, "Are spatial and global constraints really necessary for segmentation?" in *ICCV*. IEEE, 2011, pp. 9–16.
[35] A. Martinović, M. Mathias, J. Weissenberg, and L. Van Gool, "A three-layered approach to facade parsing," in *ECCV*, 2012, pp. 416–429.
[36] G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database." *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.
[37] P. Sturgess, K. Alahari, L. Ladicky, and P. H. S. Torr, "Combining appearance and structure from motion features for road scene understanding." in *BMVC*, 2009, pp. 1–11.
[38] C. Zhang, L. Wang, and R. Yang, "Semantic segmentation of urban scenes using dense depth maps," in *ECCV*, 2010, pp. 708–721.
[39] G. Floros, K. Rematas, and B. Leibe, "Multi-class image labeling with top-down segmentation and generalized robust p^n potentials," in *BMVC*, 2011, pp. 1–11.
[40] L. Ladicky, P. Sturgess, K. Alahari, C. Russell, and P. H. S. Torr, "What, where and how many? combining object detectors and crfs," in *ECCV*, 2010, pp. 424–437.
[41] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *CVPR*, 2013, pp. 1–8.

Carlo Gatta obtained the degree in Electronic Engineering in 2001 from the Università degli Studi di Brescia (Italy). In 2006 he received the Ph.D. in Computer Science at the Università degli Studi di Milano (Italy), with a thesis on perceptually based color image processing. In September 2007 he joined the Computer Vision Center at Universitat Autónoma de Barcelona as a postdoc researcher working mainly on medical imaging. He is member of the Computer Vision Center and the BCN Perceptual Computing Lab. He is currently a senior researcher at the Computer Vision Center, under the Ramon y Cajal program. His main research interests are image processing, medical imaging, computer vision, machine learning, contextual learning and unsupervised deep learning.

Francesco Ciompi received the Master's degree in Electronic Engineering from the Università di Pisa in July 2006 and the Master's degree in Computer Vision and Artificial Intelligence from Universitat Autònoma de Barcelona in September 2008. In July 2012 he obtained the PhD (cum laude) in Applied Mathematics and Analysis at the Universitat de Barcelona, with a thesis on machine learning for intravascular ultrasound images analysis. In February 2013 he joined the Universitat Autònoma de Barcelona as postdoctoral researcher, working on machine learning for computer vision and large scale image classification and retrieval. From September 2007 he is also member of the Computer Vision Center. His research interests include techniques of computer vision, machine learning and contextual learning applied to segmentation, classification and detection problems in medical imaging and pattern recognition.