Weighted atlas auto-context with application to multiple organ segmentation

Amaral, Telmo ; Kyriazakis, Ilias ; McKenna, Stephen; Plotz, Thomas

Published in:
2016 IEEE Winter Conference on Applications of Computer Vision (WACV)

DOI:
10.1109/WACV.2016.7477605

Publication date:
2016

Document Version
Peer reviewed version

Citation for published version (APA):
Amaral, T., Kyriazakis, I., McKenna, S. J., & Plotz, T. (2016). Weighted atlas auto-context with application to multiple organ segmentation. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE. DOI: 10.1109/WACV.2016.7477605

Weighted atlas auto-context with application to multiple organ segmentation

Telmo Amaral
Open Lab, School of Computing Science, Newcastle University, Newcastle upon Tyne, UK

Ilias Kyriazakis
School of Agriculture, Food and Rural Development, Newcastle University, Newcastle upon Tyne, UK

Stephen J. McKenna
CVIP, Computing, School of Science and Engineering, University of Dundee, Dundee, UK

Thomas Plötz
Open Lab, School of Computing Science, Newcastle University, Newcastle upon Tyne, UK
thomas.ploetz@newcastle.ac.uk

Abstract

Difficulties can arise from the segmentation of three-dimensional objects formed by multiple non-rigid parts represented in two-dimensional images. Problems involving parts whose spatial arrangement is subject to weak restrictions, and whose appearance and form change across images, can be particularly challenging. Segmentation methods that take into account spatial context information have addressed these types of problem, which often involve image data of a multi-modal nature. An attractive feature of the auto-context (AC) technique is that a prior "atlas", typically obtained by averaging multiple label maps created by experts, can be used as an initial source of contextual data. However, a prior obtained in this way is likely to hide the inherent multi-modality of the data. We propose a modification of AC in which a probabilistic atlas of part locations is iteratively improved and made available as an additional source of information. We illustrate this technique with the problem of segmenting individual organs in images of pig offal, reporting statistically significant improvements in relation to both conventional AC and a state-of-the-art technique based on conditional random fields.

1. Introduction

Challenges can arise from the segmentation of three-dimensional objects formed by multiple non-rigid parts represented in two-dimensional images. Biomedical image analysis problems, in particular, can involve parts whose spatial arrangement is subject to weak restrictions, and whose appearance and form change across images (e.g. depending on the presence of pathologies). Limited control over camera viewpoints can lead to occlusions between parts, and parts may sometimes be missing altogether. Segmentation methods that incorporate spatial context information have dealt with these difficulties, through the combination of inference techniques such as belief propagation (BP) [22] with models like conditional random fields (CRFs) [11]. Nevertheless, such methods typically involve complex training procedures and integrate context by means of constrained and inflexible spatial neighbourhoods. These lim-

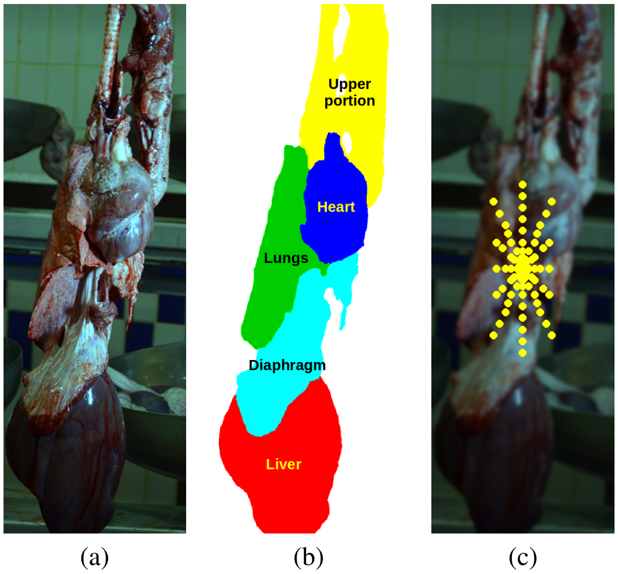


Figure 1. (a) Example image of pig "pluck" and (b) corresponding manual annotation. (c) 91-point context "stencil" overlaid on example image. (Best viewed in colour.)

itations are partially addressed by auto-context (AC) [20], an iterative technique that combines local appearance with context data based on broad and flexible spatial neighbourhoods. AC has been applied to a wide variety of previous biomedical imaging problems [20, 14].

A prior "atlas" of constituent parts can be easily obtained by rigidly registering and averaging label maps created by domain experts. Such an atlas can be used in the initial iteration of AC. The above-mentioned problems associated with segmentation of non-rigid objects, however, translate into a population of images that is multi-modal. This makes it difficult to obtain a good prior representation of part locations by simple averaging, as useful mode-specific information will become lost in the process. There is extensive literature dealing with the construction of unbiased atlases for multi-modal data, especially in the area of brain magnetic resonance (MR) image analysis, as in the work of Blezek and Miller [3] and Zikic et al. [24]. Some related work makes use of the AC model. Kim et al. [9], for example, employed an approach similar to that of Zikic et al., training multiple models, each based on an individual annotated image, so that the probability map of a new image was obtained by averaging the maps predicted by each individual model. Zhang et al. [23] proposed a clustering approach to the creation of a hierarchy of AC models (whose bottom level is similar to the set of models used by Zikic et al. and Kim et al.). Given a new image, only the best models in the hierarchy are selected to contribute to the final probability map. Model training via these techniques can be computationally very intensive, particularly when hundreds of annotated images are available.

In this paper we propose an approach that takes advantage of the iterative nature of AC, so that, at the end of each iteration, an atlas representation is updated to become closer to the image being used for training or testing. This improved atlas can then be used at the next iteration as an additional source of information (together with the label probability maps output by the classifier). Even when hundreds of training annotations are used, the computation of weighted atlases has a small impact on training and test times. To the best of our knowledge, the extension of AC with adaptive atlases used as an additional source of information has not been investigated before.

We applied our extension of the AC technique to the segmentation of multiple pig organs (namely the heart, lungs, diaphragm and liver) in images of non-digestive tract offal captured at abattoir. Figures 1(a) and 1(b) show one such image and its associated ground truth labelling of organs. Images of pig offal pose challenging problems characteristic of multi-modal data, such as partially or totally missing organs, occlusions, severe deformation and lack of control over viewpoint. The technique we propose achieved statistically significant improvements in performance over conventional AC across several iterations and outperformed a state-of-the-art segmentation technique based on CRFs.

In summary, the main contributions of this paper are: a) a proposed extension of the AC technique through the use of iteratively updated atlases, to render AC more suitable to the segmentation of objects whose two-dimensional representations are multi-modal; and b) its application to the task of segmenting multiple organs in pig offal, towards an automated solution for carcass inspection. Our extension of AC is potentially applicable to other problems involving the segmentation of non-rigid objects in image data of a multi-modal nature.

2. Auto-context with weighted atlases

2.1. Auto-context

Auto-context (AC) is an iterative pixel classification technique introduced by Tu and Bai [20]. At each iteration, local appearance features are combined with context features extracted from the class probability maps output at the previous iteration. A formal description of AC follows, using a notation similar to that of Tu and Bai [20].

Let S be a set of m training images X_j paired with their ground truth label maps Y_j,

$$S = \{(Y_j, X_j),\, j = 1..m\}. \tag{1}$$

At each iteration t we want to train a classifier that outputs the probability distribution $p_{ji}^{(t)}$ over labels $y_{ji} \in \{1..K\}$ for pixel i in image X_j, given image patch $X_j(N_i)$ and label probability map $P_j^{(t-1)}(i)$,

$$p_{ji}^{(t)} = p(y_{ji} | X_j(N_i),\, P_j^{(t-1)}(i)). \tag{2}$$

In $X_j(N_i)$, N_i denotes all pixels in the image patch, and $P_j^{(t-1)}(i)$ is map $P_j^{(t-1)}$ output for image X_j at previous iteration $t-1$, but now centred on pixel i.

The AC training procedure outputs a sequence of classifiers, one per iteration, and is formally described in Algorithm 1. Before the first iteration, all probability maps $P_j^{(0)}$ can be initialised using a prior atlas $Q^{(0)}$, obtained by averaging the m training label maps Y_j,

$$Q^{(0)} = \frac{1}{m} \sum_j Y_j. \tag{3}$$

At each iteration, given pixel i in image X_j, the actual feature vector input to the classifier is composed of local image features extracted from patch $X_j(N_i)$ concatenated with context features extracted from re-centered label probability map $P_j^{(t-1)}(i)$. In step 1 of the algorithm, n denotes the number of pixels in each image.

Algorithm 1 Training of conventional auto-context (AC) model.

```
Given training set $S = \{(Y_j, X_j),\ j = 1..m\}$, obtain prior atlas $Q^{(0)}$ from label maps $Y_j$ and use it to initialise probability maps $P_j^{(0)}$. For iteration $t = 1..T$:
  1. Build a training set for the iteration,
     $S^{(t)} = \{(y_{ji},\ (X_j(N_i),\ P_j^{(t-1)}(i)),\ j = 1..m,\ i = 1..n)\}$
  2. Train a classifier on image features extracted from $X_j(N_i)$ and context features extracted from $P_j^{(t-1)}(i)$.
  3. Use the classifier to obtain new probability maps $P_j^{(t)}(i)$.
```

Algorithm 2 Training of proposed weighted atlas auto-context (WAAC) model. The highlighted portions are specific to WAAC, i.e. they extend the original AC method by incorporating weighted atlases.

```
Given training set $S = \{(Y_j, X_j),\ j = 1..m\}$, obtain prior atlas $Q^{(0)}$ from label maps $Y_j$ and use it to initialise probability maps $P_j^{(0)}$. For iteration $t = 1..T$:
  1. Build a training set for the iteration,
     $S^{(t)} = \{(y_{ji},\ (X_j(N_i),\ P_j^{(t-1)}(i),\ Q_j^{(t-1)}(i)),\ j = 1..m,\ i = 1..n)\}$
  2. Train a classifier on image features extracted from $X_j(N_i)$ and context features extracted from $P_j^{(t-1)}(i)$ and $Q_j^{(t-1)}(i)$.
  3. Use the classifier to obtain new probability maps $P_j^{(t)}(i)$.
  4. Obtain updated atlases $Q_j^{(t)}(i)$ from new probability maps $P_j^{(t)}(i)$ and label maps $Y_j$.
```

Context features are the probabilities extracted from selected locations on map $P_j^{(t-1)}(i)$, including the central location that corresponds to current image pixel i. Selected locations are typically defined by a sparse star-shaped "stencil" such as that shown in Figure 1(c). Context data can be enhanced by including integral features, such as the sum of the label probabilities in the row to which the current point belongs, or the label probabilities summed over the whole image.

2.2. Extension with weighted atlases

At the end of each training iteration t, for each image X_j we can select the training annotations Y_k closest to probability map $P_j^{(t)}$ output by the classifier, assign a weight to each selected annotation, and combine them to obtain a weighted atlas $Q_j^{(t)}$,

$$Q_j^{(t)} = \frac{1}{\sum_{k \neq j} s_{kj}^{(t)} w_{kj}^{(t)}} \sum_{k \neq j} s_{kj}^{(t)} w_{kj}^{(t)} Y_k. \tag{4}$$

In Equation (4), weight $w_{kj}^{(t)}$ is a measure of similarity between label map Y_k and probability map $P_j^{(t)}$ (such as an F-score or a Rand index) and $s_{kj}^{(t)}$ is a selection variable defined as:

$$s_{kj}^{(t)} = \begin{cases} 1 & \text{if } k \in K_j^{(t)} \\ 0 & \text{otherwise} \end{cases}. \tag{5}$$

In Equation (5), $K_j^{(t)}$ denotes the set of indices of the m_w largest weights in $\{w_{lj}^{(t)} | l = 1..m\}$. Thus, m_w is a parameter of the proposed approach.

For the similarity measure $w_{kj}^{(t)}$ we chose to use the mean class F-score between label map Y_k and probability map $P_j^{(t)}$, given that in our application high precision and high recall are equally desirable. The F-score for a given class is defined as the harmonic mean of precision p and recall r for that class, that is, $2pr/(p + r)$. For each class, a high precision means that most of the predicted region is contained in the true region, whereas a high recall means that the predicted region contains most of the true region. Thus, a high F-score will normally correspond to predicted regions whose boundaries closely match those of the true regions. This is particularly important when segmenting multiple adjacent parts belonging to different classes.

At the start of a WAAC training iteration, features are extracted from the weighted atlas computed at the end of the previous iteration, in addition to conventional AC features. The training procedure for weighted atlas auto-context (WAAC) is formally described in Algorithm 2, whose highlighted portions correspond to the differences in relation to conventional AC. The first iteration can in principle be run as conventional AC, to avoid providing duplicate features to the classifier. (Note that, for any given image X_j, both $P_j^{(0)}$ and $Q_j^{(0)}$ would merely be copies of prior atlas $Q^{(0)}$.)

The diagram in Figure 2 represents the use of a trained WAAC model on a test image X (where image index j is omitted for simplicity). Red and green arrows correspond to the use of a classifier at each iteration to classify a pixel (represented by the small black square), whereas blue arrows correspond to use of Equation (4) at each iteration to obtain a weighted atlas. The large red square represents an image patch centred on the pixel being classified, used for the extraction of local appearance features.

We emphasise that WAAC uses the same number and spatial arrangement of context points as AC; in other words, there is no additional spatial context. At each iteration, WAAC combines information from two sources that are very different in nature: the probability maps output by the classifier (as in AC); and a weighted atlas obtained from the ground-truth component of training data. The WAAC training algorithm is not restricted to any particular type of classifier.

3. Experimental validation

3.1. Application domain

The quality and safety of meat products relies on visual inspection of carcasses as a means of detecting public health hazards and sub-clinical diseases. Carcass inspection also provides helpful information that can be fed back to farmers. However, effective and detailed screening to meet the requirements of health schemes is limited by the fact that manual assessment is inherently subjective and puts a strain on human resources. These limitations, together with new regulations of the European Food Safety Authority towards minimising carcass handling at abattoir [7], encourage the development of automated inspection systems.

Most existing literature on segmentation of multiple organs deals with localisation of human abdominal organs in computer tomography (CT) images, through varied techniques such as level set optimisation [10] and hierarchical atlases combined with statistical shape models [17]. Existing work associated with farming and meat inspection deals with comparatively simpler problems, typically involving the segmentation of whole groups of organs (without distinguishing them) and the estimation of proportions of muscle, fat and bone from CT images, both *in vivo* and *post-mortem* [16, 4]. The little available work on segmentation of multiple animal organs from video or photography usually aims to discern a particular organ of interest from nearby organs. Tao et al. [19] segmented poultry spleen from surrounding viscera, as an aid to automated detection of splenomegaly from ultraviolet and colour images. More recently, Jørgensen et al. [8] segmented gallbladders in chicken livers from images acquired at two visible wavelengths. Stommel et al. [18] envisaged a system for robotic sorting of ovine offal that would involve automatic recognition of multiple organs.

We have recently applied classic AC to the segmentation of multiple pig organs in images captured at abattoir, to assess the impact of complementing conventional context information with integral context features [1]. In this paper, we compare the proposed WAAC technique with both classic AC and a state-of-the-art CRF based pixel labelling technique. A robust organ segmentation method is useful as an intermediate stage in a wider system aimed at on-site screening for sub-clinical conditions, given that signs of such conditions are typically organ-specific. In this context, even modest improvements in organ segmentation performance in relation to available techniques can be of great importance, as regions assigned to the wrong organ may ultimately lead to missed or falsely detected pathologies.

3.2. Data

Our experiments were based on 350 images acquired at an abattoir for pigs, each showing a portion of pig offal known as the “pluck”. The pluck hangs from a hook and is composed of inter-connected organs that belong to the pig’s non-digestive tract, primarily the heart, lungs, diaphragm and liver. All images had 3646×1256 colour pixels and were obtained using a single lens reflex camera and LED lighting, both mounted on tripods. For each image, we had a ground truth label map identifying the regions covered by each of the four organs of interest, with a fifth class label being used to mark the upper region of the pluck, which usually contains the trachea and tongue and is of no interest for our application. An example of pluck and its ground truth labels are shown in Figures 1(a) and 1(b). The label map of the pluck depicts the upper portion, heart, lungs, diaphragm and liver in yellow, blue, green, cyan and red, respectively. Additional examples can be found in Figure 5.

On each image, the pluck had already been segmented from the background, through a relatively trivial segmentation step based on focus and hue information. Therefore, our task consisted in segmenting foreground pixels into the five classes defined above.

3.3. Local appearance features

We used local appearance features based on a multi-level Haar wavelet decomposition [13]. Each image was converted to the CIELUV colour space [12] and, for each component (L, u and v), the approximation wavelet coefficients as well as the horizontal, vertical and diagonal squared detail coefficients were obtained at three levels of decomposition. This resulted in 36 feature maps (3 colour components × 4 wavelet coefficients × 3 levels of decomposition), which were rescaled to match the original dimensions of the image.

We then sub-sampled each feature map and each label map by a factor of 20 along both dimensions. This resulted in 180×60 points per map, which was found to provide sufficient detail for our purposes.

For each point, we thus had a vector of 36 feature values together with a class label. As explained in Section 3.2, we were only concerned with points that fell within the foreground region of each image (that is, within the pluck). On

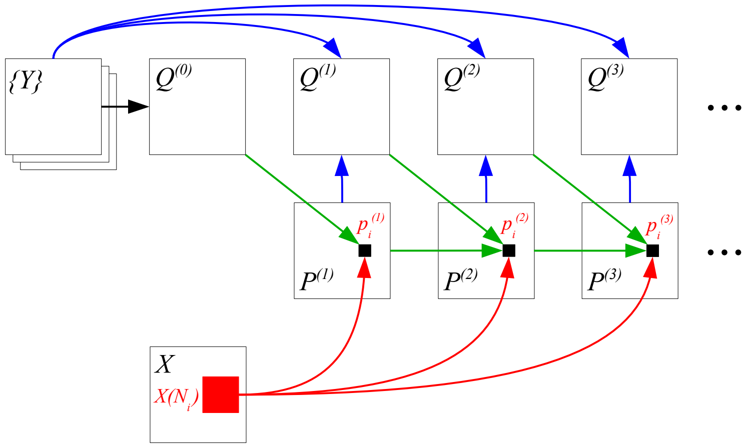


Figure 2. Diagram representing the use of a trained WAAC model. (Best viewed in colour.)

average, approximately 5,700 points per image belonged to the foreground.

3.4. Context and weighted atlas features

For each point, context features were extracted at 90 surrounding points and at the point itself. Figure 1(c) illustrates the 91 context points associated with the central point of an example image, by means of yellow markers. At the first iteration, the features extracted for each point consisted of the 5 label probabilities provided by the prior atlas at each of the 91 associated context points; at the second and subsequent iterations, they consisted of the label probabilities output by the classifier at the previous iteration, at the same context points. In addition, we complemented these features with two integral features, namely: the sum of the label probabilities in the row to which the point belonged; and the label probabilities summed over the whole image to which the point belonged. Thus, in total, for each point we extracted (91+2)×5=465 context features.

In the case of WAAC, at the second and subsequent iterations, features were extracted from the adaptive weighted atlas obtained at the end of the previous iteration, in addition to the AC context features, as explained in Section 2.2. This resulted in a total of 465×2=930 context features.

3.5. Cross-validation

We divided the 350 available images into 10 randomly picked subsets of 35 images and performed 10-fold cross-validation experiments on those subsets, to compare the performances of conventional AC and the proposed WAAC method. Thus, each cross-validation fold involved 315 training images and 35 test images.

On each cross-validation fold, the available training samples consisted of the feature vectors and class labels associated with all foreground points in the 315 training images. As each image had on average 5,700 foreground points, approximately 315×5,700=1,800,000 training samples were available per fold. From these, we obtained a balanced set of 8000 training samples, by randomly picking 1,600 samples from each of the five classes.

The training samples collected on each fold were used to train AC and WAAC classification models using Algorithms 1 and 2. Each trained model was formed by a series of multi-layer perceptrons (MLPs), one per iteration. Our MLPs had a softmax output layer and a single layer of hidden units with logistic activation. MLPs were trained to minimise regularised error $e_r = e + A\sum w^2$, where e represents the cross-entropy training error and $A\sum w^2$ is a regularisation term to prevent network weights w from becoming too large [2]. Scaled conjugate gradients optimisation was used. After experimenting with 3-fold cross-validation on a subset of training data, we opted to use 20 hidden units and a value of 0.1 for A. The implementation of MLPs provided by the NETLAB library for Matlab [15] was used.

We performed an additional 10-fold experiment to enable comparison of AC and WAAC with the state-of-the-art CRF based technique proposed by Domke [6]. For this, a 180×60 pairwise 4-connected grid was created to match the dimensions of our feature and label maps. Each model was then trained with five iterations of tree-reweighted belief propagation to fit the clique logistic loss, using a truncated fitting strategy. We used the CRF toolbox for Matlab / C++ made available by Domke [5].

On each cross-validation fold, test samples consisted of the feature vectors extracted from all foreground points in the 35 test images. Thus, approximately 35×5,700= 200,000 test samples were available per fold, all of which were used to test the trained classification models.

We studied how the overall segmentation performance (as measured by the adjusted Rand Index) varied with the number of image annotations selected for the computation

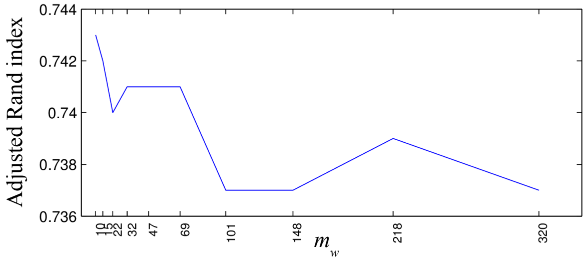


Figure 3. Variation of segmentation performance with the number of image annotations used to compute each weighted atlas (m_w).

of each weighted atlas, represented by the m_w variable implicit in Equation (4). The obtained results, shown in Figure 3, suggest that performance varied little with changes in m_w, especially taking into account that the plotted values are averages over ten folds, associated with relatively large standard deviations (≥0.027). Nevertheless, there was a tendency for better performance with low values of m_w and we set its value at 32, equivalent to 10% of the images available for training at each fold.

The adjusted Rand index is a measure of similarity between data clusterings, corrected for chance. It can be used to assess the quality of segmentation results by regarding true pixel labels and predicted pixel labels as different clusterings [21]. Defining n_{ij} as the number of pixels of class i predicted as class j, a_i as the total number of pixels of class i, and b_j as the total number of pixels predicted as class j, the adjusted Rand index can be computed using Equation (6).

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}} \qquad (6)$$

3.6. Results and discussion

Table 1 shows the average values computed over all images for two performance metrics, namely the mean class F-score and the adjusted Rand index. Results are shown for 5 iterations of AC and WAAC. Paired two-tailed Student's t-tests between AC and WAAC results yielded p-values ≤0.030 and ≤0.012 for mean class F-score and adjusted Rand index, respectively, at iterations 2 to 5.

At iteration 1, WAAC does not perform any better than AC because only the prior (non-adapted) atlas is available at that point. The evolution of each method's performance then followed the typical pattern reported by Tu and Bai [20], in that the largest improvement occurred at the 2nd iteration and performance practically levelled off by the 5th iteration. Figure 4 plots the evolution of the adjusted Rand index averaged over images, from iterations 2 to 5, with AC and WAAC.

Method	Iteration				
	1	2	3	4	5
AC	0.761	0.838	0.856	0.864	0.867
WAAC	0.756	0.847	0.862	0.868	0.871

(a) Mean class F-score

Method	Iteration				
	1	2	3	4	5
AC	0.629	0.753	0.781	0.795	0.801
WAAC	0.620	0.764	0.789	0.800	0.805

(b) Adjusted Rand index

Table 1. Average values, computed over images, of (a) mean class F-score and (b) adjusted Rand index.

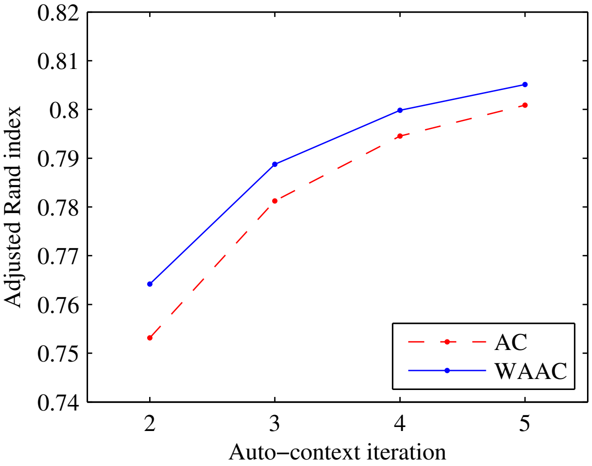


Figure 4. Evolution of adjusted Rand index (averaged over images) from 2nd to 5th iteration.

Table 2 shows the class-specific F-scores and the mean class F-score obtained with AC and WAAC, averaged over images after the 5th iteration. The largest improvement in F-score (0.012 on average) was observed for the heart. Being relatively small, the heart is the organ whose two-dimensional projection on each image is most affected by the orientation of the pluck around its vertical axis: it can be fully visible near the centre, partially or fully visible on either side of the pluck, or completely hidden. Thus, it is not surprising that the ability of WAAC to deal with multi-modality had a larger impact on the segmentation performance associated with this organ.

For three test images, Figure 5 shows the ground truth label map and the segmentation results obtained with AC and WAAC after the 5th iteration, along with the weighted atlas used by WAAC at that iteration. The general effect of WAAC is to improve spatial coherence, yielding better defined boundaries between organs, as exemplified in Figure 5(a). The weighted atlas can change dramatically depending on the image being tested, as illustrated by com-

Method	Class-specific F-score					Mean class F-score
	Upper	Heart	Lungs	Diaph.	Liver	
CRF	0.808	0.503	0.706	0.717	0.962	0.763
AC	0.928	0.744	0.855	0.821	0.955	0.867
WAAC	0.925	0.756	0.863	0.830	0.957	0.871

Table 2. Final average values, computed over images, of class-wise F-scores and mean class F-score. All improvements registered by WAAC in relation to AC are statistically significant.

Step		AC	WAAC
Once	Obtain local features	**7.206**	**7.206**
Every iteration	Obtain weighted atlas	0.000	0.516
	Obtain context features	0.114	0.153
	Normalise data	0.020	0.042
	Forward propagate MLP	0.008	0.019
	Total per iteration	**0.141**	**0.730**

Table 3. Comparison of average times needed to process a new image using trained AC and WAAC models (in seconds).

paring Figures 5(a) and 5(b). The atlas in 5(b), in particular, adapted itself to exclude the heart, allowing a much better final result than with conventional AC. Nevertheless, situations such as that shown in Figure 5(c) can occur: when a class that is normally present happens to be missing (in this case, the liver), the training annotations selected to form the weighted atlas may turn out to be substantially different from the image being tested. In the example shown, this led to the heart being almost excluded from the final result.

Table 2 also shows results obtained using Domke's CRF based segmentation method, after five iterations of tree-reweighted belief propagation. It can be seen that, after its 5th iteration, the CRF method yielded a mean class F-score merely at the level of the first iteration of AC or WAAC. The shortcomings of the CRF method when applied to our task are also clearly visible in the segmentation maps shown in Figure 5. The most obvious problem with CRF based results is that portions of organs are detected in regions where they make no sense, as happens for example with small regions of upper offal and liver segmented near the top of the diaphragm in Figure 5(a), or a small region of heart wrongly segmented within the diaphragm in Figure 5(b).

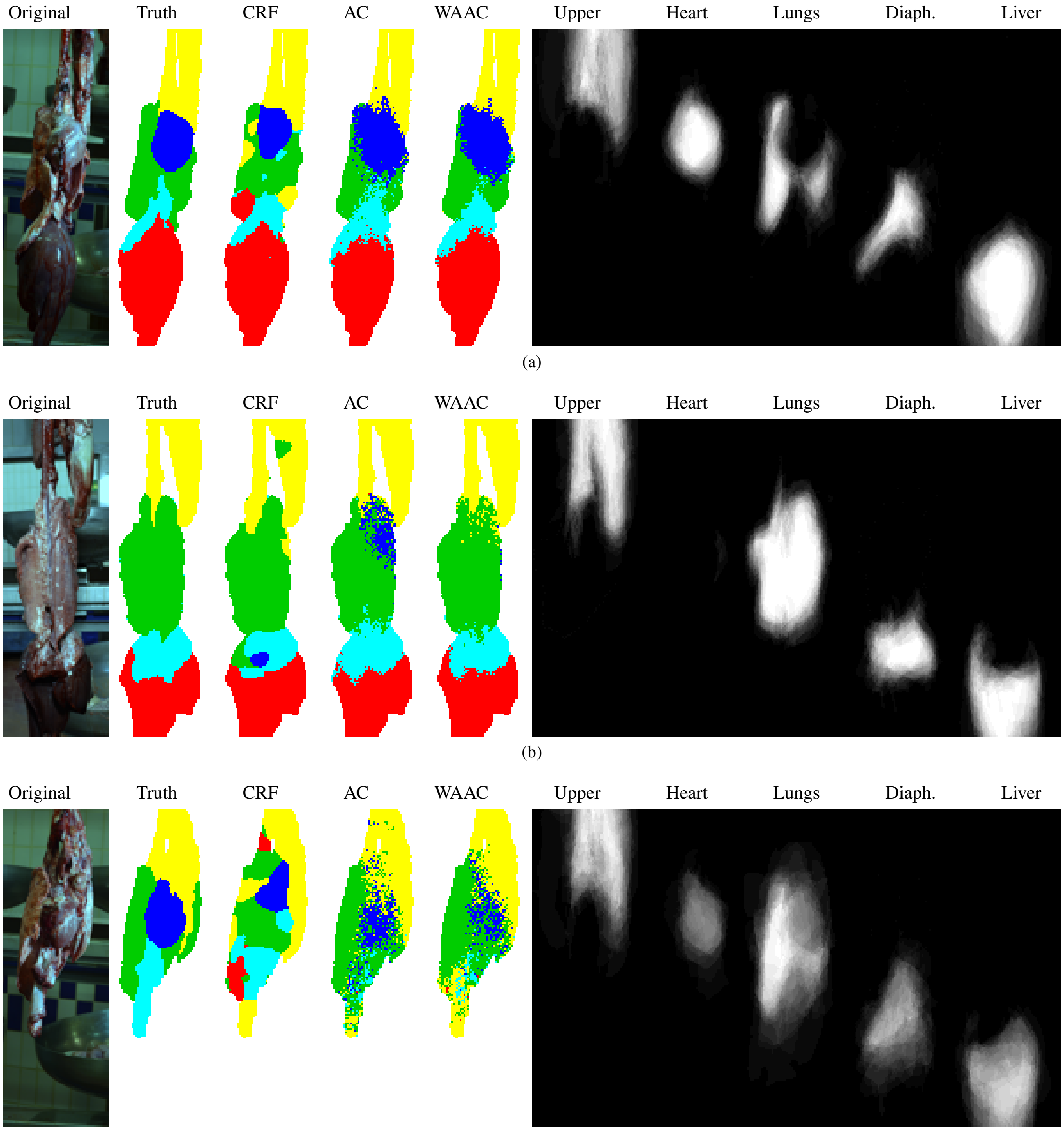
These results show that this CRF-based method is inappropriate for segmentation tasks like the one we addressed in this work, involving parts whose spatial arrangement, appearance and form vary widely across images. In contrast, taking advantage of the iterative nature of AC, the WAAC technique we propose is able to identify the training label maps that are most relevant for a given test image and use that knowledge to steer the segmentation process, thus helping to avoid the erroneous localisation of parts within conflicting contexts.

Table 3 shows the average times taken to process a new image using AC and WAAC. These times are dominated by the extraction of local appearance features, which need to be computed once for any new image, taking about 7 seconds. Then, each iteration of WAAC takes about 0.6 seconds longer than AC to process an image, mainly due to the computation of the weighted atlas. For example, classic AC took on average $7.21{+}0.14{\times}5{=}7.91$ seconds to complete 5 iterations, whereas WAAC took $7.21{+}0.73{\times}5 = 10.86$ seconds (that is, about 1.37 times longer than AC). Times were measured on a regular desktop machine, using only the CPU (an Intel Core i7-870). Our feature extraction and atlas computation routines were implemented in Matlab. The computation of weighted atlases, in particular, was optimised via the use of vectorised operations, making it easily adaptable for faster execution on GPU.

4. Conclusion

We proposed a modification of the auto-context technique in which a probabilistic atlas of part locations is iteratively improved and made available for the extraction of features, to complement those used in the conventional approach. We tested our technique on the problem of segmenting multiple organs in images of pig offal acquired in an industrial setting. Results on a large data set of images were reported, showing statistically significant improvements in segmentation performance over the traditional auto-context approach. We also demonstrated the superiority of auto-context and our proposed method over a state-of-the-art technique based on conditional random fields, applied to our problem.

Future directions of work could include the computation of weighted atlases in a class-wise fashion, the use of alternative similarity measures in the computation of the atlases, as well as the evaluation of our technique on other segmentation problems involving data of a multi-modal nature. The proposed method could be applicable to other problems involving the segmentation of non-rigid objects into their constituent parts, such as anatomical structures in medical images of various modalities, or sub-cellular compartments in microscopy images.



(a)

(b)

(c)

Figure 5. For each of three images: original image, ground truth labels, and segmentation maps obtained with CRFs, AC, and WAAC. Also shown for each image are the five components of the weighted atlas used at the 5th WAAC iteration. (Best viewed in colour.)

5. Acknowledgements

This project was co-funded by Innovate UK, the Biotechnology and Biological Sciences Research Council (BBSRC) and Tulip Ltd. It was carried out in collaboration with Tulip Ltd and Hellenic Systems Ltd. We thank Sheralyn Smith, Gareth Woodward and Lazar Valkov for the patient manual annotation of hundreds of images.

References

[1] T. Amaral, I. Kyriazakis, S. J. McKenna, and T. Plötz. Segmentation of organs in pig offal using auto-context. In *International Symposium on Biomedical Imaging (ISBI)*, 2016. (In press).

[2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] D. Blezek and J. Miller. Atlas stratification. *Medical Image Analysis*, 11(5):443–457, 2007.

[4] L. Bünger, C. Glasbey, G. Simm, J. Conington, J. Macfarlane, K. McLean, K. Moore, and N. Lambe. *CT Scanning - Techniques and Applications*, chapter Use of X-ray computed tomography (CT) in UK sheep production and breeding, pages 329–348. InTech, 2011.

[5] J. Domke. Justin's Graphical Models / Conditional Random Field Toolbox. URL http://users.cecs.anu.edu.au/~jdomke/JGMT/.

[6] J. Domke. Learning graphical model parameters with approximate marginal inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:2454–2467, 2013.

[7] Food Standards Agency. Changes to pig meat inspection in June 2014. FSA consultation, March 2014.

[8] A. Jørgensen, T. Moeslund, and E. Mølvig Jensen. Detecting gallbladders in chicken livers using spectral analysis. In *Machine Vision for Animals and their Behaviour (BMVC workshop)*, pages 2.1–2.8, 2015.

[9] M. Kim, G. Wu, W. Li, L. Wang, Y-D. Son, Z-H. Cho, and D. Shen. Segmenting hippocampus from 7.0 Tesla MR images by combining multiple atlases and auto-context models. In *Machine Learning in Medical Imaging (MICCAI workshop)*, pages 100–108, 2011.

[10] T. Kohlberger, M. Sofka, J. Zhang, N. Birkbeck, J. Wetzl, J. Kaftan, J. Declerck, and S. Zhou. Automatic multi-organ segmentation using learning-based segmentation and level set optimization. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 338–345, 2011.

[11] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1150–1157, 2003.

[12] M. Mahy, L. Eycken, and A. Oosterlinck. Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV. *Color Research & Application*, 19(2):105–121, 1994.

[13] S. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

[14] S. J. McKenna, T. Amaral, S. Akbar, L. Jordan, and A. Thompson. Immunohistochemical analysis of breast tissue microarray images using contextual classifiers. *Journal of Pathology Informatics*, 4:13, 2013.

[15] I. Nabney. *NETLAB: algorithms for pattern recognition*. Springer, 2002.

[16] E. Navajas, C. Glasbey, K. McLean, A. Fisher, A. Charteris, N. Lambe, L. Bünger, and G. Simm. In vivo measurements of muscle volume by automatic image analysis of spiral computed tomography scans. *Animal Science*, 82(04):545–553, 2006.

[17] T. Okada, K. Yokota, M. Hori, M. Nakamoto, H. Nakamura, and Y. Sato. Construction of hierarchical multi-organ statistical atlases and their application to multi-organ segmentation from CT images. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 502–509, 2008.

[18] M. Stommel, W. Xu, P. Lim, and B. Kadmiry. Robotic sorting of ovine offal: Discussion of a soft peristaltic approach. *Soft Robotics*, 1(4):246–254, 2014.

[19] Y. Tao, J. Shao, K. Skeeles, Y. Chen, et al. Detection of splenomegaly in poultry carcasses by UV and color imaging. *Transactions of the ASAE-American Society of Agricultural Engineers*, 43(2):469–474, 2000.

[20] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(10):1744–1757, 2010.

[21] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward objective evaluation of image segmentation algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):929–944, 2007.

[22] J. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. In *Neural Information Processing Systems (NIPS)*, volume 13, pages 689–695, 2000.

[23] L. Zhang, Q. Wang, Y. Gao, G. Wu, and D. Shen. Learning of atlas forest hierarchy for automatic labeling of MR brain images. In *Machine Learning in Medical Imaging (MICCAI workshop)*, pages 323–330. Springer, 2014.

[24] D. Zikic, B. Glocker, and A. Criminisi. Atlas encoding by randomized forests for efficient label propagation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 66–73, 2013.