Exploiting Sparsity for Real Time Video Labelling

Lachlan Horne, Jose M. Alvarez, and Nick Barnes
College of Engineering and Computer Science, Australian National University, and
NICTA Canberra Research Laboratory
Tower A, 7 London Circuit, Canberra ACT 2600, Australia.
{lachlan.horne, jose.alvarez, nick.barnes}@nicta.com.au

Abstract

Until recently, inference on fully connected graphs of pixel labels for scene understanding has been computationally expensive, so fast methods have focussed on neighbour connections and unary computation. However, with efficient CRF methods for inference on fully connected graphs, the opportunity exists for exploring other approaches. In this paper, we present a fast approach that calculates unary labels sparsely and relies on inference on fully connected graphs for label propagation. This reduces the unary computation which is now the most computationally expensive component. On a standard road scene dataset (CamVid), we show that accuarcy remains high when less than 0.15 percent of unary potentials are used. This achieves a reduction in computation by a factor of more than 750, with only small losses on global accuracy. This facilitates real-time processing on standard hardware that produces almost state-of-the-art results.

1. Introduction

Real-time video labelling is important for computer vision applications such as autonomous driving and driver assistive technologies, but current methods are computationally expensive and impractical for time-critical applications such as pedestrian detection.

In this paper we propose a framework to enable real-time pixel-wise semantic segmentation of video. Our method uses a frame-by-frame approach which is amenable to real-time streaming applications such as road hazard detection, and can incorporate data from earlier frames for improved accuracy in some cases. We reduce the amount of computation required by reducing the number of unary potentials which are calculated in each frame. Figure 1 shows typical outputs of our algorithm using different numbers of unary potentials per frame. We show experimentally that for single images from street scenes, accuracy remains high when only a small fraction of unary potentials are used. We achieve a factor of 768 computational saving for unary potentials on single images with negligible loss of global accuracy. We further show that our approach for video can achieve improved accuracy with the same computational saving per frame, and give a more temporally consistent result.

In Section 2 we describe work related to this paper. In Section 3 we describe our system in detail, and then in Section 4 we present experimental results for our system on a publicly available dataset. Our experiments show that our approach can reduce the amount of computation required for video labelling without significant loss of accuracy, enabling real-time video labelling.

2. Related Work

Multi-class segmentation of images and video is a challenging task, and has been the focus of much recent computer vision research. Often the problem is posed as a maximum a posteriori (MAP) inference in a conditional random field (CRF) defined over image pixels or patches [15, 7, 4]. CRF potentials incorporate local information about a pixel or patch as unary potentials, and smoothness terms to maximize label agreement between similar pixels or patches as pairwise potentials. The calculation of unary potentials can take a wide variety of image features into account [13] and having high quality unary potentials is considered critical to producing a correct labelling.

The basic CRF model only uses pairwise terms over neighbouring pixels or patches, which limits its ability to model long-range connections within the image. In order to improve segmentation and labelling accuracy, many approaches have been explored, such as hierarchical methods [8] and higher order potentials [9]. However, these methods are computationally demanding [1], and not well-suited to segmenting video in real-time.

Recent work in fully-connected CRFs is promising. Fully-connected CRFs use pairwise potentials which connect every pixel or patch in an image or video. These have been used with some success [12, 14, 5, 11] but were lim-

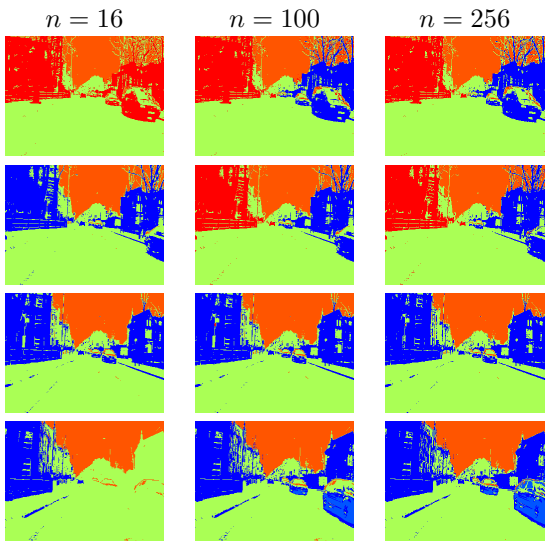


Figure 1: Outputs of our algorithm on four consecutive frames from the CamVid day test set with different unary potential densities and no interlacing. n refers to the number of unary potentials computed per frame. Higher unary densities allow for finer detail to be segmented. 256 unary potentials per frame corresponds to a factor of 300 reduction in the number of unary potentials calculated.

ited by the vast complexity of the problem when posed as a fully-connected CRF. This complexity has been overcome recently with work in efficient approximate inference on fully-connected CRFs [10], which can produce a pixel-wise labelling for an image, given precomputed unary potentials, in 0.2 seconds. However, the calculation of unary potentials still limits our ability to perform labelling with CRF models in real time video. Unary potentials must be generated for each pixel, which remains a prohibitively time consuming process for realtime applications. For example, [13] reports 1 second per image. Real time performance requires the labelling for each video frame to be completed before the next video frame is ready, thus allowing less than 100 milliseconds per frame at 10Hz. It may be feasible to complete inference in this time, but in order to achieve real-time labelling, we must reduce the time taken for unary potentials by a significant factor. This is where our contribution is focused.

3. Algorithm

Our framework, outlined in Figure 2, uses the fully-connected CRF inference method of [10] to produce a pixel-wise multi-class labelling of input frames. Our contribution is in the generation of unary potentials.

Our unary potential generation comprises a *unary selection* method, which determines at which pixel locations

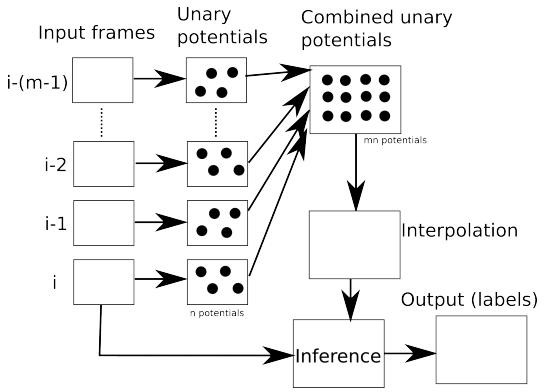


Figure 2: Our framework for video, showing data dependencies for the i^{th} frame. Note that only unary potentials for the previous m frames are used, and in fact previous frames can be discarded with only the unary potentials retained in memory.

unary potentials shall be calculated, and what should be done with pixel locations which do not have unary potentials calculated. We extend this to video by incorporating a *temporal interlacing* method, which exploits expected correlation between video frames by using unary potentials calculated for previous frames.

Pairwise potentials for the CRF inference are generated from pixel locations and color values. Thus for most pixel locations, due to the sparsity of calculated unary potentials, only the CRF encodes local information. This means that the CRF is critical for accurately calculating positions and shapes of image regions.

The modularity of this framework makes it feasible to test a variety of approaches for unary selection and temporal interlacing, and future work should investigate this. In this paper we focus on results on single images using a simple grid-based unary selection method with bilinear spatial interpolation, and results on video using interlacing over four frames.

3.1. Unary Selection

The locations at which unary potentials are calculated can affect the quality of the labelling produced. Figure 3 shows the effect of sparse unary selection when small features are present in the image. It is possible to "miss" a small image region and not calculate unary potentials within it, which is likely to lead to a mislabelling of the region. Changing the locations can change which regions are missed, so overall accuracy can vary significantly. Similarly, when unary potentials vary over a region (due to noise or clutter in the image) point selection has a strong effect, which would not be the case when using dense unary potentials, since the effects of noise would more likely be averaged out. Mislabelling can happen when the selected points

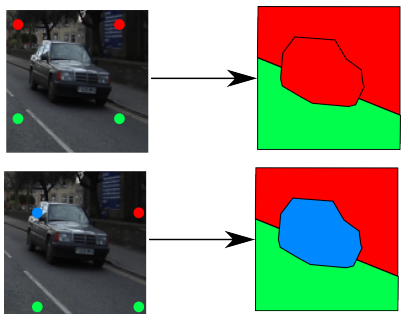


Figure 3: An example of the strong effect of unary selection. In this artificial example, two different unary selections are shown. Unary potential locations are shown as circles, colored according to the most likely label for that point. Even with perfect classifiers, if no unary potentials are calculated within the region of the car, that region is labelled incorrectly. But due to the CRF inference, even if there is only one unary potential calculated in that region, the region can be labelled correctly. This is an idealized example, but this behaviour is often observed in practice.

align with artifacts in the unary potentials.

One way to mitigate this effect is to compute unary potentials at locations which are more likely to be within regions of similar pixels. We experimented with using centroids of superpixels generated by SLIC, as this could avoid cases where small regions are missed by unary selection. However we found this did not give significant improvement over a simple static grid of locations, particularly when taking the extra computational requirements of superpixel calculation into account. It can help avoid the case where unary potentials are calculated at strong image gradients, but these are not neccessarily locations where the pixel classifier we use (based on TextonBoost) gives ambiguous results. This idea could yet be expanded upon in future work.

We generate a regular grid of locations in the image at which to calculate unary potentials. We vary the number of unary potentials calculated by varying the density of the grid. This is a simple approach with negligible computational cost. We respect the aspect ratio of the image, so that the same number of rows and columns of grid points is used. We also add an offset to the grid locations to avoid potential artifacts caused by unary potentials being calculated at the edges of the image.

All pixels require valid unary potential values for CRF inference. We attempted setting the non-computed potentials to a constant value. This means that the label of the node is only determined via inference, since changing the label of a node would result in no direct change in the energy of the graph. However we found this was unable to achieve sufficiently high accuracy with small numbers of calculated unary potentials. We found interpolation to be important for maintaining accuracy as the grid density is reduced, since it provides a strong signal for pixels far from unary potential locations.

We use bilinear interpolation for spatial interpolation of unary potentials. Since we use a regular grid of unary potential locations, this interpolation approach is straightforward to apply. It is also fast, and commonly implemented in realtime on consumer graphics processors. We interpolate the potentials for each label separately.

3.2. Interlacing

We observe that in video, there is usually high correlation between subsequent frames, particularly for cases such as vehicle-mounted cameras, where the camera egomotion is limited. We propose exploiting this by reusing unary potentials from previous frames when labelling each frame. In our system, each frame uses a different unary selection (eg. a different grid offset) such that new unary potentials calculated for each frame are in different locations from the reused unary potentials. Since camera egomotion is likely to be small between subsequent frames in the vehicle-mounted camera case, we expect that the error introduced by this approach will be small. Since we are using data from multiple frames to label each frame, we expect a temporal smoothing effect, which should improve the temporal stability of the labelling.

In this paper we use interlacing over cycles of predetermined unary selections. That is, we generate m arrangements of n image locations which are fixed throughout each experiment. When labelling each frame, we calculate unary potentials using one of these arrangements, and reuse unary potentials calculated for the previous $m-1$ frames, each with a different point arrangement. This means that at each frame, only n unary potentials are calculated, but mn unary potentials are used in labelling. This is shown diagrammatically in Figure 4. We expect this should also mitigate the effect of spatially noisy unary potentials, since the effective density of unary potentials is higher.

4. Experiments

We evaluate our approach on the Cambridge Video dataset (CamVid) [3, 2], a challenging benchmark for multi-class segmentation for road scenes. This dataset consists of video frames captured at 30Hz from a vehicle-mounted camera, driving around Cambridge in moderate traffic conditions. Hand-labelled ground truth is provided for one in 30 frames for an effective rate of 1Hz, with each pixel assigned one of 32 semantic classes. Following the experimental setup of [2], the images are down-scaled to 320×240 and the semantic classes are grouped into 11 categories.

Quantitative evaluation is performed using pixel-wise comparisons of the obtained segmentations with ground-

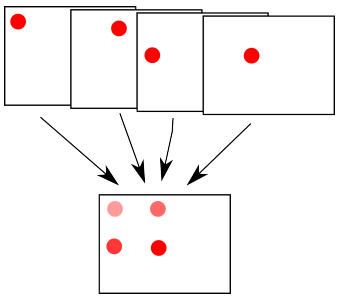


Figure 4: Interlaced unary potentials over four frames. Unary potentials are calculated at different positions in each frame, then combined when used. A repeating cycle of unary potential locations is used.

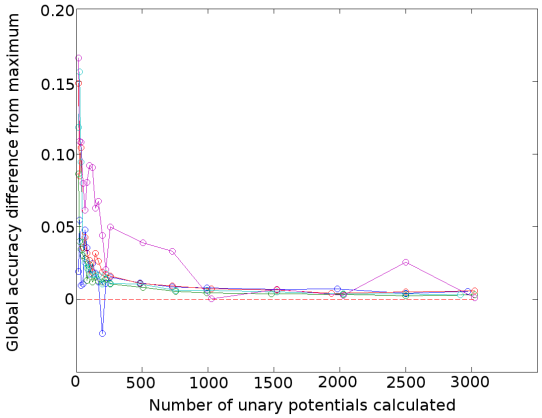


Figure 5: Global accuracy discrepancy from maximum (all unary potentials with CRF) for 5 randomly selected frames. Values for each frame are averaged over 4 trials with different unary potential grid offsets. We can observe a common trend that accuracy values rapidly approach the maximum. One outlier is shown - some frames require more unary potentials to reach high accuracy and can unexpectedly lose accuracy at certain grid densities. This is often due to small (narrow) regions being missed by coarse grids, and repeating patterns (such as *fence* regions in CamVid) being sensitive to grid densities more than offsets.

truth. We report the global and per-class average accuracies. The former represents the ratio of correctly classified pixels to the total the number of pixels in the test set. The latter is computed as the average over all classes of the ratio of correctly classified pixels in a class to the total number of pixels in that class. We also show qualitative results on video, to assess and demonstrate the effect of interlacing on temporal stability and qualitative labelling performance.

We employed the CamVid training set to learn pixel-level classifiers using the publicly available DARWIN framework of [6]. These classifiers were used to generate unary potentials on the test image set. The features used were the 17-dimensional features used in [13] incorporating local color, texture and location. Parameters for the CRF inference and pairwise term generation were kept fixed throughout all experiments.

For comparison, we perform labelling with unary potentials calculated for all pixel locations, with and without CRF inference (shown in Table 2). Using all unary potentials with CRF inference gives the maximum possible accuracy of our method, and we base our comparison on this. Our maximum accuracy is comparable to other state of the art methods.

4.1. Single Image Results

Since the position of the grid points can drastically affect the output labelling, we used multiple trials of randomly chosen grid offsets to ensure a fair comparison. We tested frames from the CamVid day test set with 4 trials per frame, averaging results over the trials. Table 1 shows the global accuracy of our method over these frames. We can see that results rapidly converge to near-maximum accuracy for most frames in the sequence. This effect does depend on image content, variation over different frames is shown in Figure 5. We can see in Table 1 that we only need around 100 calculated unary potentials to achieve similar accuracy to the maximum. This results in a factor of 768 reduction in computation time for unary potentials, which is more than sufficient for real-time performance. Increasing n can further increase accuracy, but instead we use interlacing as another method to increase accuracy without increasing computation required.

We examined the labellings produced with small numbers of unary potentials generated and in most cases they did not differ greatly from the maximum accuracy labellings. We show a typical example in Figure 6. Smaller regions can be missed by unary potential point selection, and be labelled differently due to unary potentials being interpolated. In this case however, the smaller regions were segmented from the surrounding region due to noise, and in fact in such cases it is possible for accuracy to be higher than that for the case where all unary potentials are calculated.

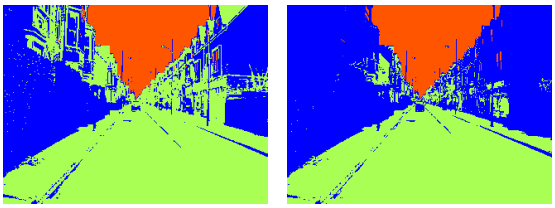
4.2. Video Results

We performed experiments to test unary potential interleaving on CamVid. For these experiments we fixed the number of unary potentials calculated per frame to 100, since the single image experiments show that the accuracy for most frames is close to maximum by this point, but some improvement is still possible.

We test video labelling with and without interlacing. We do not test interlacing with more than four frames at a time, since interlacing over too many frames constrains our ap-

n	Proportion	Unary (interpolated)	Pairwise
64	0.08%	12.6	74.9
81	0.11%	11.9	75.3
100	**0.13%**	**11.3**	**76.1**
144	0.19%	10.5	76.9
196	0.26%	9.82	77.4
225	0.29%	9.57	77.5
256	0.33%	9.31	77.7
484	0.63%	8.49	78.7
729	0.95%	7.97	79.1
1024	1.33%	7.69	79.5
2025	2.64%	7.31	80.0
76800	100%	80.5	80.8

Table 1: Global accuracy mean as a function of the number of unary potentials calculated (n) over frames from the CamVid day test set. The Proportion column shows the proportion of image pixels with calculated unary potentials. There is a clear advantage to using pairwise inference when sparse unary potentials are used. Accuracy with unary potentials falls as the number of unary potentials increases due to the effect of small regions being picked up and having their labels propagated by interpolation. Pairwise inference mitigates this effect by taking image content into account when propagating labels. We highlight values for $n = 100$ since we use this value of n in later experiments.



(a) (b)

Figure 6: Qualitative results for single-image labelling. Note that smaller features may not be segmented from large regions, as expected. (a) Labels using unary potentials for all pixels and pairwise inference. (b) Labels using 100 unary potentials, bilinear interpolation and pairwise inference.

proach to video with small motion. We expect that in the CamVid dataset, motion over four subsequent frames is small enough that it is unlikely to introduce errors.

We specifically want to compare two methods over video:

No interlacing Only the unary potentials of the current frame are used. The same grid locations are used in every frame.

4-cycle interlacing Four grids are used, each with a different offset. Unary potentials from the past 3 frames are used for each frame.

n	100	100	76800
Interlacing	None	4	None
Building	92.6	94.5	94.3
Car	25.4	42.0	72.5
Column_Pole	0.05	0.10	5.08
Fence	0.67	4.95	5.43
Pedestrian	0.00	0.22	6.60
Road	98.6	98.3	97.8
Sidewalk	8.34	18.4	31.6
Sign_Symbol	0.04	0.84	3.52
Sky	98.1	97.9	98.1
Tree	39.6	47.4	53.1
Global accuracy	75.9	78.5	80.8

Table 2: Per-class accuracy comparison between methods over the CamVid day test set. All results are produced with fully-connected inference, with the only variation between columns being the unary potentials used. The rightmost column shows the maximum possible performance of our approach, with all unary potentials calculated. n refers to the number of unary potentials calculated per frame and indicates the amount of computation required per frame - which is reduced by a factor of 768 where $n = 100$. In the interlaced case $m = 4$ so there are effectively 400 unary potentials used per frame. Note that interlacing increases global accuracy as expected. Note also that interlacing improves performance slightly on most labels, and significantly for Building, Sidewalk and Fence. Bicyclist was not present in the day test set and so is not shown in this table.

Note that neither of these methods are informed by image content or domain-specific information. It may be possible to improve performance by increasing grid density and decreasing the number of frames used at once, thus improving spatial and temporal precision, in regions where precise labelling is more important, such as in peripheral and central regions for pedestrian detection in autonomous vehicle applications.

We tested the above methods on frames with ground truth from the CamVid day test set. We used earlier frames as inputs for interlacing. Since the video in CamVid was captured at 30Hz, 4-cycle interlacing will only use data from a temporal window of 133 milliseconds.

The accuracy results in Table 2 show the interlacing approach does improve labelling accuracy as expected, and performs better for some labels than others. In particular we observe improvement for classes which typically manifest as small or narrow image regions, such as Car, Fence or Sidewalk. This suggests that the increased spatial density of unary potentials allows detection of these classes. Notably

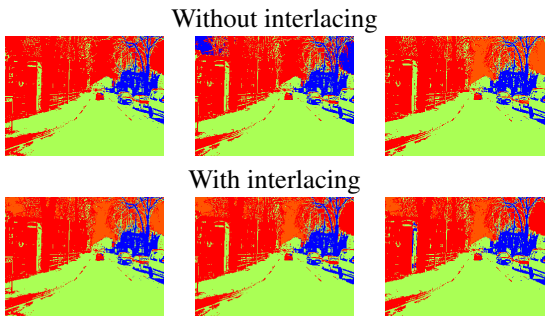


Figure 7: Qualitative results for interlaced versus non-interlaced video labelling. The improved temporal stability can be seen in the top-left and top-right corners of frame, with large regions showing less sudden label variation between subsequent frames. Some instability is still present with interlacing and this could be improved by increasing density of calculated unary potentials (increasing calculation time per frame) or increasing the number of frames used for interlacing (making the system less robust to strong motion).

performance for Pedestrian and Sign_Symbol remains poor, which is most likely due to these regions still being small relative to the denser grid. This suggests future work could characterise the relationship between object size and grid density, and find a method to determine the optimal unary potential density for detecting certain object classes.

In order to confirm that the expected temporal smoothing effect of unary interlacing does in fact improve temporal stability, we compare frames which exhibit some temporal artifacts in the non-interlaced cases (ie. labels "flickering") and observe the difference when the unary potentials are interlaced. Figure 7 illustrates typical results with and without interlacing.

5. Conclusions

We proposed a method to enable real-time semantic labelling of video, using sparse generation of unary potentials and dense CRF inference to reduce the overall computation required. We achieved this with negligible loss of accuracy. Further, we propsed an interlacing approach to improve quantitative and qualitative performance on video without significant increase in computational requirements. In doing this we achieved significant accuracy gains for certain object classes.

Acknowledgments

NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council (ARC) through the ICT Centre of Excellence program. This research was supported by the ARC through its Special Research Initiative (SRI) in Bionic Vision Science and Technology grant to Bionic Vision Australia (BVA).

References

[1] K. Alahari, P. Kohli, and P. Torr. Dynamic hybrid algorithms for discrete map mrf inference. In *IEEE Trans. PAMI*, 2010. 4322

[2] G. J. Brostow, J. Fauqueur, and R. Cipolla. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters*, xx(x):xx–xx, 2008. 4323

[3] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. Segmentation and recognition using structure from motion point clouds. In *ECCV*, 2008. 4323

[4] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009. 4322

[5] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, 2008. 4322

[6] S. Gould. Darwin: a framework for machine learning and computer vision research and development. *J. Mach. Learn. Res.*, 13(1):3533–3537, Dec. 2012. 4324

[7] S. Gould, J. Rodgers, D. Cohen, G. Elidan, and D. Koller. Multi-class segmentation with relative location prior. *IJCV*, 80(3):300–316, 2008. 4322

[8] X. He, R. S. Zemel, and M. . Carreira-perpin. Multiscale conditional random fields for image labeling. In *CVPR*, 2004. 4322

[9] P. Kohli and P. H. S. Torr. Robust higher order potentials for enforcing label consistency. In *CVPR*, 2008. 4322

[10] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. *CoRR*, abs/1210.5644, 2012. 4322

[11] N. Payet and S. Todorovic. (rf) 2 random forest random field. In *Proc. NIPS*, 2010. 4322

[12] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007. 4322

[13] 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. *IJCV*, 81(1):2–23, Jan. 2009. 4322, 4324

[14] T. Toyoda and O. Hasegawa. Random field model for integration of local information and global information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1483–1489, Aug. 2008. 4322

[15] J. Verbeek and B. Triggs. Scene segmentation with crfs learned from partially labeled images. In *NIPS*, 2008. 4322