Label Propagation in Video Sequences

Vijay Badrinarayanan[†] Fabio Galasso[†] and Roberto Cipolla
University of Cambridge, UK
{vb292,fg257,cipolla}@cam.ac.uk

Abstract

This paper proposes a probabilistic graphical model for the problem of propagating labels in video sequences, also termed the label propagation problem. Given a limited amount of hand labelled pixels, typically the start and end frames of a chunk of video, an EM based algorithm propagates labels through the rest of the frames of the video sequence. As a result, the user obtains pixelwise labelled video sequences along with the class probabilities at each pixel. Our novel algorithm provides an essential tool to reduce tedious hand labelling of video sequences, thus producing copious amounts of useable ground truth data. A novel application of this algorithm is in semi-supervised learning of discriminative classifiers for video segmentation and scene parsing.

The label propagation scheme can be based on pixelwise correspondences obtained from motion estimation, image patch based similarities as seen in epitomic models or even the more recent, semantically consistent hierarchical regions. We compare the abilities of each of these variants, both via quantitative and qualitative studies against ground truth data. We then report studies on a state of the art Random forest classifier based video segmentation scheme, trained using fully ground truth data and with data obtained from label propagation. The results of this study strongly support and encourage the use of the proposed label propagation algorithm.

1. Introduction

The problem of label propagation has received some attention from the machine learning community for the task of semi-supervised learning using both labelled and unlabelled data points [14]. On the other hand researchers in computer vision have paid only marginal attention to this important problem, addressing it as "label transfer" across similar images in a database [9] or recognising objects (by labelling corresponding pixels) using a trained image generative model [7], both requiring large quantities of labelled training images. In any case to the best of our knowledge this problem has not been considered for video sequence data, possibly because it appears as a (deceptively) simple task given the strong correlation between successive frames. In response, our contributions are two fold:

1. We propose a probabilistic framework and algorithm for label propagation in video sequences;
2. We demonstrate effective semi-supervised learning with label propagation for video segmentation and recognition on new challenging sequences [11, 4].

Our primary interest in this work is to transfer labels (road, pedestrians, cars and the like) from the two labelled ends of a video sequence to the remaining unlabelled frames (See Fig. 1 for an illustration). Additionally, we are also interested in obtaining a measure of how confident a label assigned to a pixel is, or in familiar terms, the class distribution at the pixel labels. In Section 3 we develop a probabilistic model to incorporate these requirements and perform maximum likelihood based inference.

The probabilistic model proposed in this paper, under different settings, allows one to compare naive methods for label propagation using optic flow estimates, for instance, alongside more sophisticated approaches based on image patches

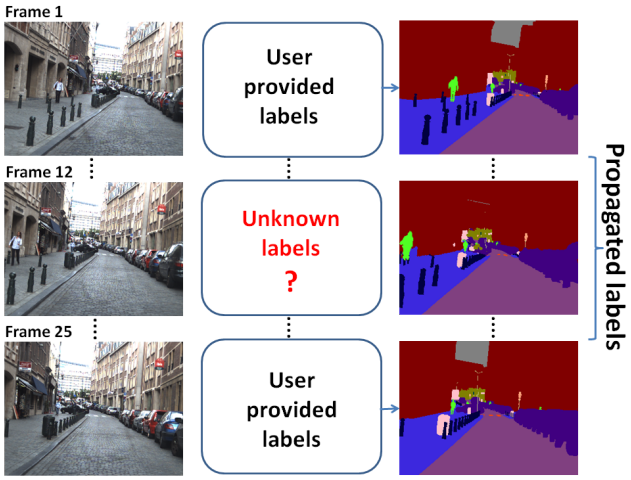

Figure 1. An illustration of label propagation in video sequences

[†] indicates equal contribution

[5] or extraction of semantically consistent regions [1]. The first set of experiments in Section 4 perform a comparative study of these methods and demonstrates both quantitatively and qualitatively that the proposed label propagation methods are superior to naive solutions under various experimental settings and ground truth. The second set of experiments in Section 4 aims to convince the reader that the propagated labels from the proposed algorithms can indeed be used to train state of art discriminative classifiers like Random Forests [11] for video sequence classification with minimal loss in accuracy of classification. We conclude and discuss future prospects in Section 5.
The following section presents a literature review.

2. Related Work

A well known method for label propagation in machine learning literature is that of Zhu *et al.* [14]. They formulate their label propagation problem as a problem of assigning soft labels to nodes of a fully connected graph with few labelled nodes. Weights are assigned to the links according to the proximity of the corresponding nodes in the graph and an iterative update is proposed for propagating labels to the unlabelled nodes. This method is shown to perform well in assigning labels to a hand written digits database. Other such methods, based on exploiting the "geometry of the data", are also available in the machine learning literature (see [3] for a detailed survey).
In contrast, sequential data and their labelling are more naturally modelled using directed graphs. Particularly, we draw inspiration from the "static" epitome model for videos in [5]. We extend this to a "coupled" Hidden Markov Model (HMM) which can employ pixels, image patches or semantic regions. There is an implicit "inpainting" idea behind these models which is particularly useful for video sequence labelling where occlusions (disocclusions) are frequent.
Label propagation has also been proposed in the context of multi-label learning [6], where data points are assigned their class label under the learnt influence of the correlation between the class labels. Although meaningful, these methods need copious amounts of labelled training data, which is difficult to gather in the first place. Instead, we avoid any sort of training for label propagation and rely on a generative model to perform accurate labelling.
Only recently in the computer vision community some research has been devoted to label transfer from training images in a set of similar (and labelled) images in a database [9] to a given test image. In particular, the authors formulate their "image matching" method in the SIFT-descriptor space using the traditional optic flow optimization setup. Hence, they argue that their method is the equivalent of optic flow correspondence for non-sequential data. However, some surprising and flawed matches are obtained, inconsistent with the quantitative energy term they evaluate. In this work, we demonstrate by comparisons that optic flow based labelling is less efficient than the proposed methods for label propagation.
Elsewhere, [7] demonstrate the ability of their trained generative Jigsaw model for images to transfer labels from a learnt jigsaw (source) to an image in the training (labelled) dataset. Similarly discriminative models like CRFs [8] also strive to perform joint segmentation and (pixelwise) recognition, but require a large dataset of carefully labelled (similar) images. At this juncture, straightforward extensions of such models for video sequences are not available and for that purpose, one can definitely anticipate the need for large amounts of labelled (video) data. The proposed algorithms of this paper are exactly suited to this purpose.
As a video labelling tool, the proposed algorithms can be contrasted with currently operative annotation tools. A representative sample of these is LabelMe Video [12]. Here the user draws a (rough) polygon around an object at the start, some key frames and the end frame, and defines a rich set of annotations (category, static or moving, occluded). Labelling is achieved via interpolation with a 3D motion model, while tags are used to learn the category priors and estimate the 3D structure of the scene. We base our "tool" on a generative model which jointly explains the observed frames and their annotations, with the resulting capability to obtain both pixelwise labels and their probabilities. We employ these results to train discriminative classifiers for pixelwise recognition.

3. Label Propagation

The proposed graphical model (see Fig. 2) is a coupled HMM for the joint generative modelling of image sequences (continuous variables) and their annotation (discrete variables). While [5] learn a "static" epitome model with space-time patches for video modelling (inpainting & video interpolation are their target applications), we employ a time-series model for video annotation and avoid learning a video epitome. However, in our model we incorporate their key idea of inducing correlations between image patches in the inference stage. The elements of our model are described below.

IMAGE MODELLING Shaded node variables $I_{0:n}$ represent the observed sequence of images. Hidden variable Z_k can represent a set of mutually independent colour image pixels, rectangular image patches or even semantic regions [1] (each colour channel is treated independently). For clarity's sake we only describe the case of rectangular image patches here (See [2] for details for the other two cases).
Conceptually, I_{k-1} predicts the set of latent patches Z_k which in turn are used to explain (generate) observation I_k

and so on along the Markov chain.

$\mathbf{I_{k-1} \rightarrow Z_k}$: following [5] this link is defined as follows.

$$p(Z_k|I_{k-1}, T_k) = \prod_{j=1}^{\Omega_k} \prod_{i \in j} \mathcal{N}(z_{k,j,i}; I_{k-1,\mathcal{T}_{k,j}(i)}, \Phi_{k-1,\mathcal{T}_{k,j}(i)}) \tag{1}$$

where, index j runs over all the (overlapping) latent patches $Z_k = \{Z_{k,j}\}_{j=1}^{\Omega_k}$. $z_{k,j,i}$ is pixel i inside patch j at time k. Each patch j has an associated variable $T_{k,j}$ which indexes (overlapping) patches $\{1, \ldots \Omega_k\}$ in I_{k-1}. $T_k = \{T_{k,j}\}_{j=1}^{\Omega_k}$ is the collection of patches. Note that the number, size and ordering of patches are the same in Z_k and I_{k-1}. Following this, $\mathcal{T}_{k,j}(i)$ indexes the corresponding pixel $I_{k-1,\mathcal{T}_{k,j}(i)}$ in patch $I_{k-1,\mathcal{T}_{k,j}}$.

$\mathcal{N}(z_{k,j,i}; I_{k-1,\mathcal{T}_{k,j}(i)}, \Phi_{k-1,\mathcal{T}_{k,j}(i)})$ is a normalized Gaussian distribution over $z_{k,j,i}$, with mean $I_{k-1,\mathcal{T}_{k,j}(i)}$ and variance $\Phi_{k-1,\mathcal{T}_{k,j}(i)}$. This distribution quantifies the ability of image patches in I_k to predict latent patches in Z_k.

$\mathbf{Z_k \rightarrow I_k}$: this link is defined as follows.

$$p(I_k|Z_k) = \prod_{v \in V} \mathcal{N}(I_{k,v}; \frac{1}{N_v} \sum_{\substack{j=1 \\ s.t. j \supset v}}^{\Omega_k} z_{k,j,v}, \Psi_{k,v}), \tag{2}$$

where $I_{k,v}$ denotes the intensity of pixel v in the pixel grid V. j indexes patches in Z_k which overlap pixel v. $\Psi_{k,v}$ is the variance of the normalized Gaussian distribution.

ANNOTATION MODELLING Let $l = 1 \ldots L$ index the different object classes, where $l = 1$ corresponds to an unlabelled (void) class (See [2] for more details on void class). Hidden variable A_k is an image sized grid. $\{a_{k,v,l}\}_{l=1}^{L}$ are a set of positive real valued parameters at pixel v of this grid, which obey $\sum_{l=1}^{L} a_{k,v,l} = 1$. Thus, $\{a_{k,v,l}\}_{l=1}^{L}$ represent the class distribution for the corresponding pixel in I_k. The end variables of the chain, A_0, A_n (shaded) are initialized using Eqn. 12.

Hidden variable Z_k^a can represent a set of mutually independent "annotated" image pixels, rectangular image patches or semantic regions. Adjacent to each Z_k^a on the chain are variables A_{k-1} and A_k. As in the image model, A_{k-1} predicts the set of annotated patches Z_k^a, which in turn is used to predict A_k and so on along the bottom Markov chain.

$\mathbf{A_{k-1} \rightarrow Z_k^a}$: this link is defined as follows.

$$p(Z_k^a|A_{k-1}, T_k) = \prod_{j=1}^{\Omega_k} \prod_{i \in j} \prod_{l=1}^{L} a_{k-1,\mathcal{T}_{k,j}(i),l}^{z^a_{k,j,i,l}}, \tag{3}$$

where the indices on the first two products are the same as in Eqn.1. The last term is the discrete class probability distribution of the pixel $z^a_{k,j,i}$ in patch $Z^a_{k,j}$.

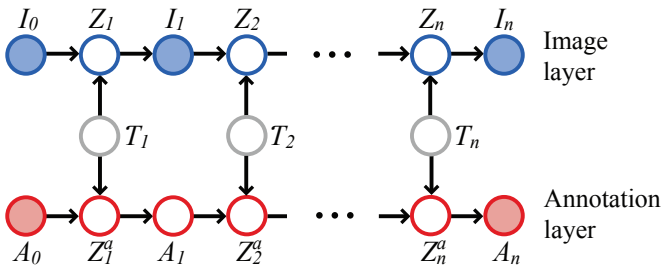

Figure 2. Proposed graphical model for label propagation.

$\mathbf{Z_k^a \rightarrow A_k}$: this link is defined as follows.

$$p(A_k|Z_k) = \prod_{v \in V} \frac{\Gamma(\alpha_{v,0})}{\Gamma(\alpha_{v,1}) \cdots \Gamma(\alpha_{v,L})} \prod_{l=1}^{L} a_{k,v,l}^{\alpha_{v,l}-1}, \tag{4}$$

which is a Dirichlet prior on the (independent) parameters $\{a_{k,v}\}_{v \in V}$. Γ denotes the gamma function with parameters $\alpha_{v,l} = \frac{1}{N_v} \sum_{\substack{j=1 \\ s.t. j \supset v}}^{\Omega_k} z^a_{k,j,v,l}$ for $l = 1 \ldots L$ and $\alpha_{v,0} = \sum_{l=1}^{L} \alpha_{v,l}$. Note that j indexes patches in Z_k^a which overlap pixel index v in the pixel grid V.

3.1. Inference

Given $\{I_{0:n}, A_0, A_n\}$, we estimate the latent variables $\Theta = \{Z_{1:n}, Z^a_{1:n}, A_{1:n-1}, \mathcal{T}_{1:n}\}$ using the variational EM algorithm. The log of the data likelihood is lower bounded as shown below.

$$\log p(I_{0:n}, A_0, A_n) \geq \int_{\Theta} q(Z_{1:n}, Z^a_{1:n}, A_{0:n}, \mathcal{T}_{1:n}) \times \log \frac{p(Z_{1:n}, Z^a_{1:n}, A_{0:n}, \mathcal{T}_{1:n}, I_{0:n})}{q(Z_{1:n}, Z^a_{1:n}, A_{0:n}, \mathcal{T}_{1:n})}. \tag{5}$$

For tractability, we assume the following form for the auxiliary distribution (See [5]).

$$q(Z_{1:n}, Z^a_{1:n}, A_{0:n}, \mathcal{T}_{1:n}) = \prod_{k=1}^{n} q(\mathcal{T}_k) \times \delta(Z_k - Z_k^*)\delta(Z_k^a - Z_k^{a*})\delta(A_k - A_k^*). \tag{6}$$

The delta terms above imply that we infer the most probable hidden states.

EXPECTATION STEP Fixing the latent variables to $A^*_{0:n}, Z^*_{1:n}, Z^{a*}_{1:n}$ (note $A_0^* = A_0, A_n^* = A_n$) , the E-step computes the posterior over the mapping variables:

$$q(\mathcal{T}_{k,j}) = p(\mathcal{T}_{k,j}|I_{0:n}, A^*_{0:n}, Z^*_{1:n}, Z^{a*}_{1:n})$$
$$\propto \prod_{i \in j} \mathcal{N}(z^*_{k,j,i}; I_{k-1,\mathcal{T}_{k,j}(i)}, \Phi_{k-1,\mathcal{T}_{k,j}(i)}) \prod_{l=1}^{L} a_{k-1,\mathcal{T}_{k,j}(i),l}^{z^{a*}_{k,j,i,l}} \times p(\mathcal{T}_{k,j}). \tag{7}$$

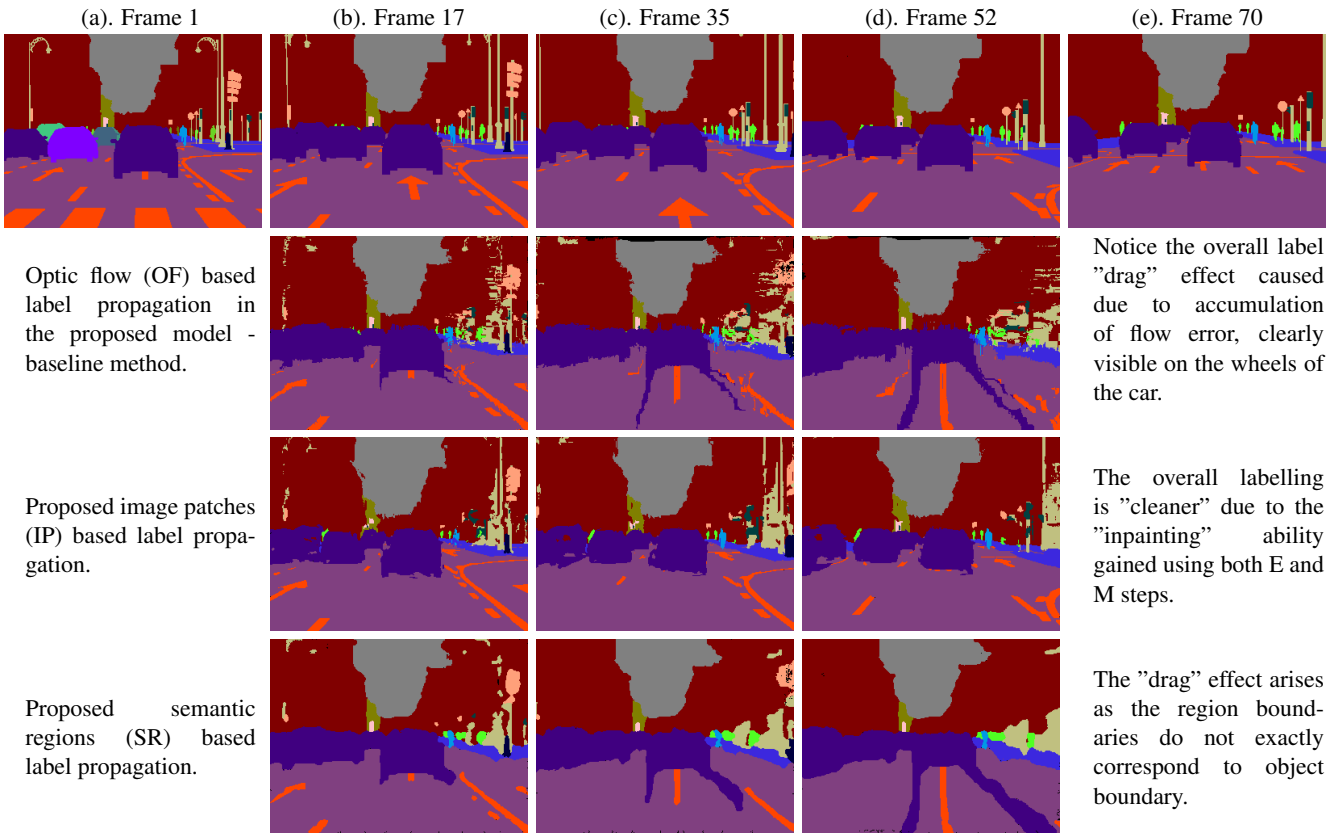

Figure 3. Seq 1 - Frames 1 & 70 on the top row are user provided labels. Ground truth for frames 17, 35 & 52 are provided for comparison on the top row. The proposed IP based method delivers the best labelling

Note that in this step the posterior of the mapping variables is dependent on both the observed image data and the current estimate of the labels and their probabilities.

MAXIMIZATION STEP The lower bound under the E-step distributions is to be maximized wrt $A^*_{1:n-1}, Z^*_{1:n}, Z^{a*}_{1:n}$. Conditioned on the mapping variables, $\{T_k\}_{k=1}^n$, the estimation of $Z^*_{1:n}$ and $A^*_{1:n-1}, Z^{a*}_{1:n}$ can be separated. This leads to the following.

$$z^*_{k,v} = \frac{\frac{I_{k,v}}{\Psi^2_{k,v}} + \sum_{j=1}^{\Omega_k} \sum_{\mathcal{T}_{k,j}} q(\mathcal{T}_{k,j}) \frac{I_{k-1,\mathcal{T}_{k,j}(v)}}{\Phi^2_{k-1,\mathcal{T}_{k,j}(v)}}}{\frac{1}{\Psi^2_{k,v}} + \sum_{j=1}^{\Omega_k} \sum_{\mathcal{T}_{k,j}} \frac{q(\mathcal{T}_{k,j})}{\Phi^2_{k-1,\mathcal{T}_{k,j}(v)}}}, \forall v \in V. \tag{8}$$

For the remaining parameters we follow an alternation strategy to obtain their estimates. We fix $A^*_{1:n}$ and optimise the lower bound to get,

$$z^{a*}_{k,v,l} = \begin{cases} 1 & \text{if} \quad \nabla z^{a*}_{k,v,l} > \nabla z^{a*}_{k,v,l'}, l' = 1, ..., L, l' \neq l, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

Using the above estimate,

$$a^*_{k,v,l} \propto \tag{10}$$
$$\underbrace{(z^{a*}_{k,v,l} - 1) + \sum_{j=1}^{\Omega_{k+1}} \sum_{\mathcal{T}_{k+1,j}} \sum_{\substack{i \\ s.t. \mathcal{T}_{k+1,j(i)} = v}} q(\mathcal{T}_{k+1,j}) z^{a*}_{k+1,j,i,l}}_{b_{k,v,l}},$$

which upon normalization delivers,

$$a^*_{k,v,l} = \frac{b_{k,v,l}}{\sum_{l=1}^{L} b_{k,v,l}}, l \in 1 : L. \tag{11}$$

Eqns. 9, 10, 11 are alternated, in that order, to obtain the estimates of the hidden variables at convergence.

INITIALIZATION The variables $Z^a_{2:n-1}$ are all initialized to zero without affecting the iterations. Z_1, Z_n are clamped to the user provided labels and are not updated throughout. The parameters A_0, A_n are initialized as follows.

$$a_{k,v,l} = \begin{cases} 1 & \text{if user provided class label is l,} \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

The variances of the normalized Gaussians in Eqns. 1, 2 are fixed to 5.0.

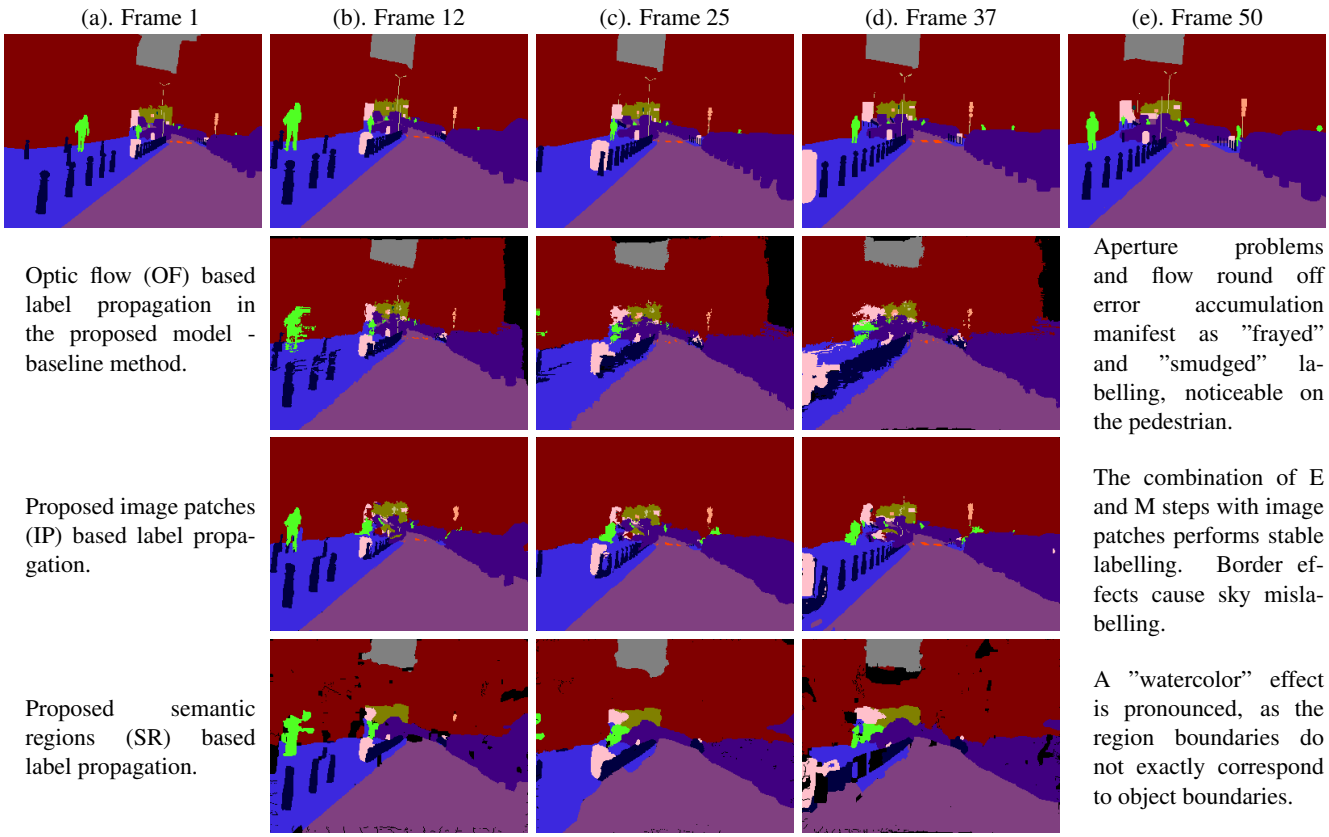

Figure 4. Seq 2 - Frames 1 & 50 on the top row are user provided labels for label propagation. Ground truth for frames 12, 25 & 37 are provided for comparison on the top row. The proposed IP method performs best

EVIDENCE PROPAGATION In this work we propagate the labels in two full passes over all the hidden variables. Further details on the relation between the variational approximation in Eqn 6 and its effects on label propagation can be found in [2].

4. Experiments

The first experiment compares the accuracy of propagated labels against known ground truth, using: pixel wise correspondences as delivered by an optic flow (OF) algorithm [13], rectangular image patch (IP) matches [5], and pre-optimised mappings of semantically consistent regions (SR) [1] (see Section 3). The results are relevant to object cutout in videos, extracting masks for alpha matting for cinema post-production and other interactive applications. Next, we study the test effects of training a state of the art Random Forest classifier [11] using the results of label propagation under different settings. However, due to space constraints we only report the results of training the classifier using the IP mapping strategy.

DATASET DESCRIPTION We use a new and challenging pixelwise ground truthed video dataset. This dataset consists of three outdoor driving video sequences (VGA resolution) captured using a camera placed on a car. The ground truths are available for 70 frames for Seq 3 & Seq1, and 98 frames of Seq 2. 14 different classes are labelled (sky, building, road, pavement, pedestrian, car and the like). This ground truth is obtained via tedious hand labelling, costing a minimum of about 45-60 minutes per frame.

TESTING THE ACCURACY OF LABEL PROPAGATION Using only 1 iteration with 6×6 sized image patches and 75% overlap between patches, the E-step (Eqn. 7) computes the posterior probability of the mapping variables. A "flat" prior over a 30×40 search area surrounding the center of the patch is used. In case of OF and SR based mappings, the posterior over mappings is replaced with *deterministically available approximate mappings*, pre-computed optic flow and pre-computed tracks of semantically consistent regions ([1]) obtained via dynamic programming (See Appendix for details) respectively. In particular, region track optimisation can be bracketed in spirit to semi-dense particle flow computations [10]. Both use a forward backward strategy to deliver globally "optimal trajectories" to handle occlusions (disocclusions) but region tracks also provide dense matches. Probabilistic (IP) and deterministic (OF,SR) map-

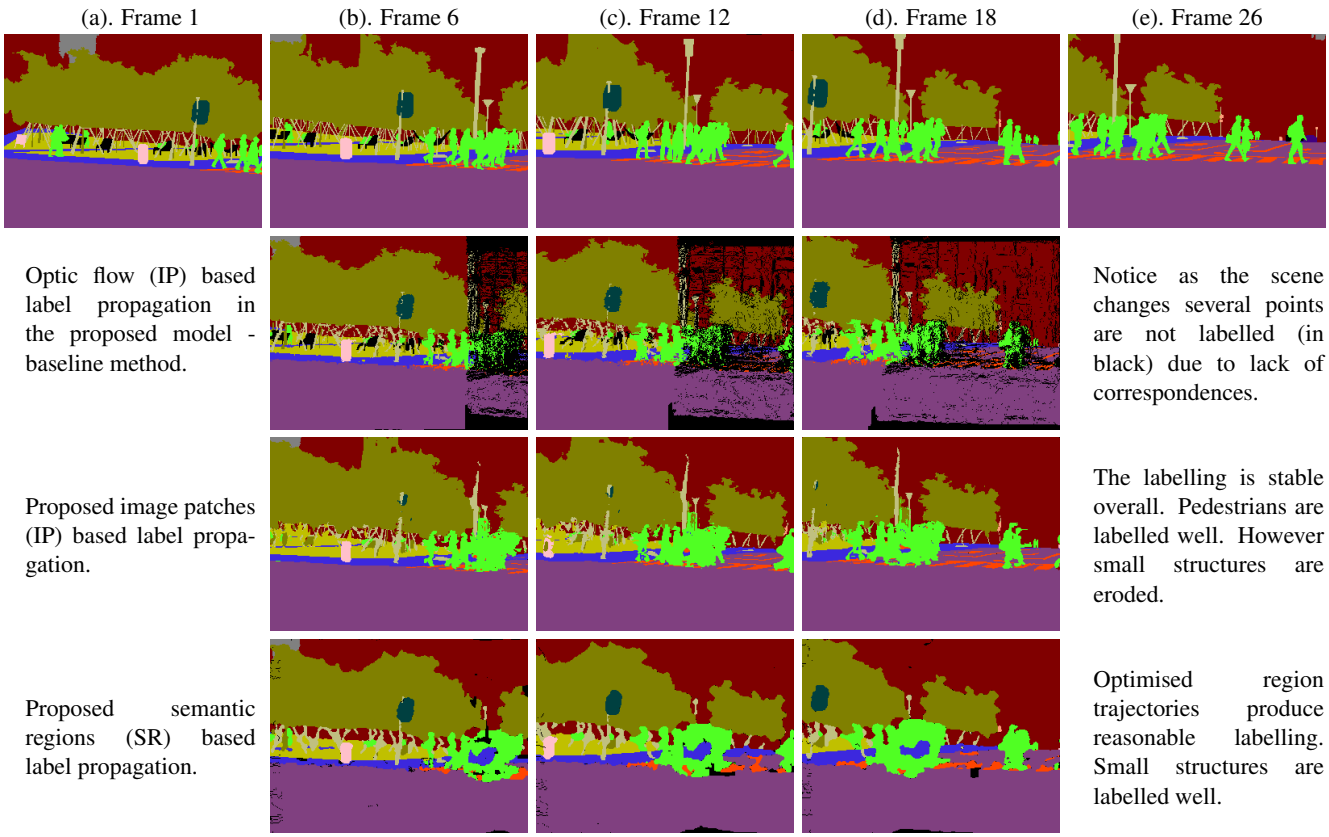

Figure 5. Seq 3 - Frames 1 & 25 on the top row are user provided labels. Ground truth for frames 6, 12 & 18 are provided for comparison on the top row. The proposed methods propagate labels fairly accurately under a panning motion here

pings are evaluated over increasing width between the two ends of the video sequences, 25, 50 & 70 frames.

TRAINING RANDOM FOREST CLASSIFIERS This experiment is a study of semi-supervised learning using label propagation. We choose the image patch based mapping strategy as the test label propagation method of choice (see Table 1) and a Random forest [11] (with 15 trees and a maximum depth of 10) as the classifier of choice. 98 frames of Seq 1 is chosen as the test sequence, as maximum ground truth is available for this sequence. Seq 1 and Seq 3 are chosen as training sequences. The classifier is trained under three different settings; under fully ground truthed Seq 1 and Seq 3, using ground truth for Seq 1 and label propagation for Seq 3, and finally, using label propagation for both Seq 1 and Seq 3. These settings are also evaluated over varying lengths of propagated labels (25, 50 and 70 frames) over the training sequences.

RESULTS AND DISCUSSIONS

Accuracy test Fig. 6 reproduces the quantitative results of the tests on Seq 1, 2 & 3. In Seq 1 (Fig. 3), image patch (IP) based mapping outperforms the other methods under occlusions (disocclusions) due to *probabilistic mapping* and its *inpainting* ability. The qualitative result on Fig. 3 vindicate these numerical results. In Seq2 (Fig. 4), the class average accuracies are highest for the OF mapping, this is primarily due to the mislabelling of the ”sky” class in IP based mapping (attributable to border effects). In contrast, the global accuracy and the qualitative result in Fig .4 clearly indicate the superiority of the IP based mapping over the other two. It is interesting to note the slower degradation rate of both global and class average accuracy for IP based mapping as the width increases as compared to the other two. Of interest is also the fact that even with a slight lower accuracy on pedestrian and car class the qualitative effect is better for IP based mapping for Seq 2. Categories such as pavements and road markings, which are very useful for driving applications, are labelled better by IP based mapping. In Seq 3 (Fig. 5), the scene changes quickly and new objects appear (disappear). Here the SR based mapping which uses *sequence optimized trajectories* demonstrates highest class average accuracy, closely followed by IP based mapping (for 25 and 51 frames). The IP based mapping degrades over the 70 frame test due to its inability to correctly explain unseen (provided in ground truth) objects. The naive OF based mapping leaves large amounts of data points unlabelled (this is not counted into the void class which is re-

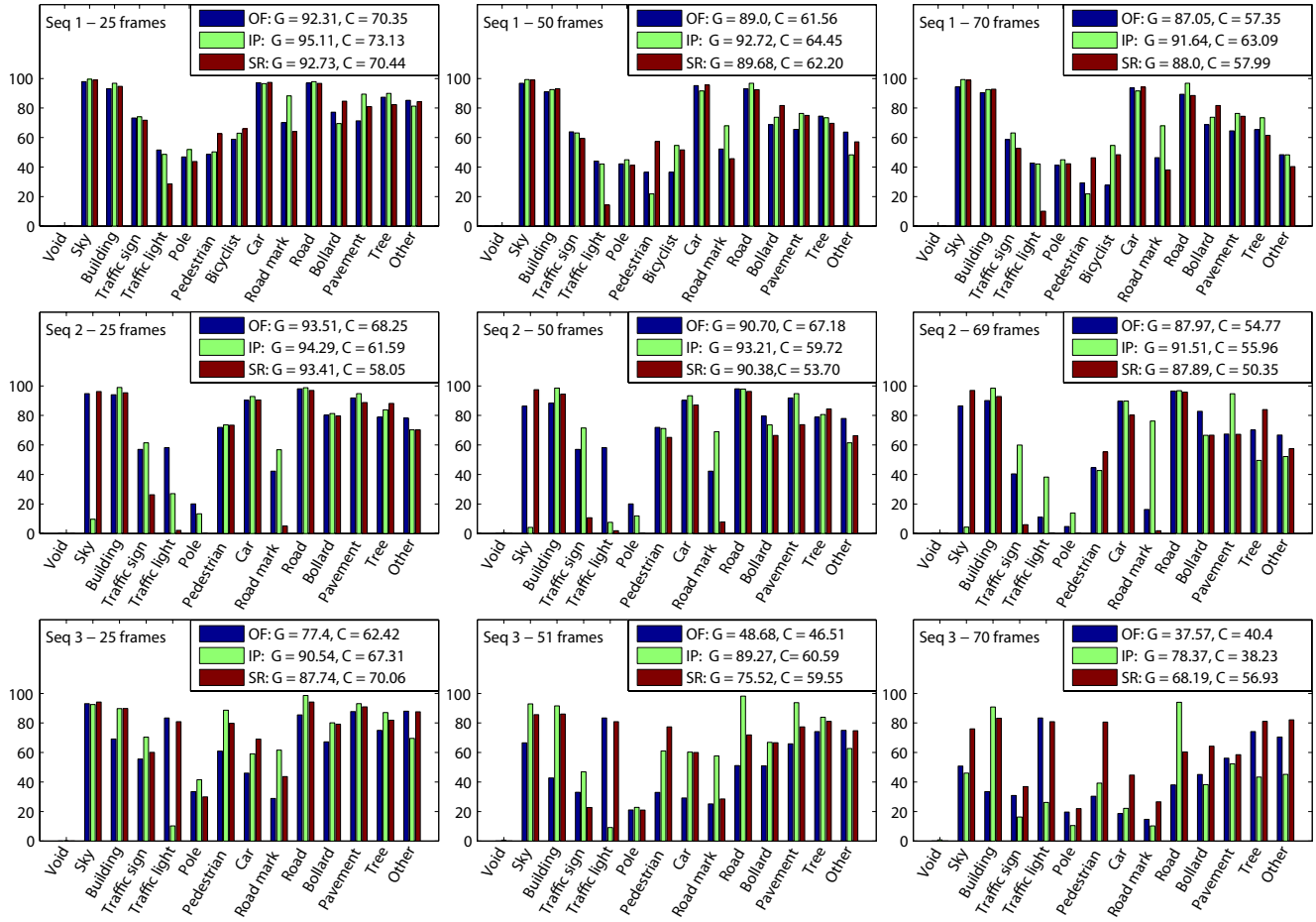

Figure 6. Quantitative comparison of OF(baseline), IP and SR label propagation methods. The reader is suggested to glance at the inset global(G) and class average(C) accuracies for a quick comparative summary

served for unlabelled pixels in the ground truth) as the camera pans. This is the cause of an apparent abberation (see the OF results for 70 frames), in that, the class average is higher than the global average. To summarise, the results in Figs. 3, 4, 5 & Fig. 6 indicate that at least one of the two proposed methods (IP and SR based mappings) is superior to optic flow based label propagation. From Fig. 6 it is clear that OF based method is outperformed in 7 out of the 9 experiments by the proposed IP or SR based approaches.

Semi supervised learning Table 1 reports results of video classification experiments with Random forests. The test accuracy (on Seq 2) undergoes little degradation (class average) as ground truth is progressively replaced by propagated labels. From the second row (50 frames) it appears that test accuracy is higher when trained with propagated labels as compared to ground truth based training, a fact which can be attributed to randomization in classifier training. Finally, as the width between end frames is increased to 70 frames, the class average accuracy when trained with propagated labels degrades only about 5%. These results should encourage training of classifiers based on the proposed methods.

5. Conclusion

We have proposed a probabilistic generative model for label propagation in video sequences. The inference mechanism propagates labels to the unlabelled parts of the video sequence in a transductive (batch) setting. Over short sequences naive optic flow based mapping under this model performs acceptable label propagation. More sophisticated probabilistic mappings using image patches or deterministic pre-computed region trajectories provide accurate label propagation even in longer and more challenging sequences. By means of both qualitative and quantitative results we have demonstrated that the proposed methods provide a tool to extract large quantities of labelled ground truth, which are useful for semi-supervised learning of discriminative classifiers.

The variational approximation in Eqn 6 yields only the most probable (MP) value of the hidden states. This approxima-

	Settings under which the Random Forest is trained		
Width	Full ground truth (Seq1 & Seq3)	Ground truth (Seq 1) + Propagated labels(Seq 3)	Propagated labels (Seq1 & Seq 3)
25 frames	G = 62.1%, C = 44.6%	G = 61.6%, C = 44.6%	G = 59.8%, C = 44.2%
50 frames	G = 60.5%, C = 45.5%	G = 60.5%, C = 47.2%	G = 60.7%, C = 46.1%
70 frames	G = 57.3%, C = 45.5%	G = 58.7%, C = 42.2%	G = 58.8% , C = 40.3%

Table 1. Test results of Random Forest based classification of Seq 2 trained under three different lengths of training sequences Seq 1 and Seq 3 and three different training settings. The comparable test accuracy to training under ground truth provides support for training classifiers using the proposed methods

tion leads to tractable inference of the MP values. The drawback is that only an "instantaneous" notion of label uncertainty can be captured based on the MP values (see Eqns 9, 10, 11). In the future we aim to introduce more complex approximating distributions to propagate label uncertainties. We also aim to extend the model to employ discriminative classifiers for label propagation in longer sequences.

Acknowledgements

The authors wish to acknowledge the helpful suggestions of Ignas Budvytis, Tae-Kyun Kim and Yu Chen. Thanks also to the funding agencies for their support.

References

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *CVPR*, 2009.

[2] V. Badrinarayanan, F. Galasso, and R. Cipolla. Label propagation in video sequences. Technical report, CUED/F-INFENG/TR-643, March 2010.

[3] Y. Bengio, O. Delalleau, and N. Le Roux. Label propagation and quadratic criterion. In *Semi-Supervised Learning*, pages 193–216. MIT Press, 2006.

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

[5] V. Cheung, B. J. Frey, and N. Jojic. Video epitomes. In *CVPR*, 2005.

[6] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, pages 1719–1726, 2006.

[7] A. Kannan, J. Winn, and C. Rother. Clustering appearance and shape by learning jigsaws. In *NIPS, Volume 19.*, 2006.

[8] P. Kohli, L. Ladicky, and P. Torr. Robust higher order potentials for enforcing label consistency. *IJCV*, 82(3):302–324, 2009.

[9] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: Label transfer via dense scene alignment. In *CVPR*, 2009.

[10] P. Sand and S. Teller. Particle video: Long-range motion estimation using point trajectories. *IJCV*, 80(1):72–91, 2008.

[11] J. Shotton, M. Johnson, and R. Cipolla. Semantic texton forests for image categorization and segmentation. In *CVPR*, 2008.

[12] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. Labelme video: building a video database with human annotations. In *ICCV, Kyoto, Japan*, 2009.

[13] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime tv-l1 optical flow. In *Pattern Recognition (Proc. DAGM)*, pages 214–223, Heidelberg, Germany, 2007.

[14] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical report, CMU-CALD-02-107, Carnegie Mellon University., 2002.

Appendix: Extraction of semantic region tracks

The optic flow $\bar{\mathbf{u}}(\mathbf{x})$ at a pixel $\mathbf{x} = (x, y)$ is estimated using the algorithm of [13] and smoothed using bilateral filtering (as in [10]) to improve boundary sharpness:

$$\bar{\mathbf{u}}(\mathbf{x}) = \frac{\Sigma_{\mathbf{x}' \in r_f} \bar{\mathbf{u}}(\mathbf{x}') w(\mathbf{x}, \mathbf{x}')}{\Sigma_{\mathbf{x}' \in r_f} w(\mathbf{x}, \mathbf{x}')} , \tag{13}$$

where r_f is the region [1] at frame f and the function $w(\mathbf{x}, \mathbf{x}')$ weighs the neighbouring pixels $\mathbf{x}'$ of $\mathbf{x}$ according to spatial proximity and motion similarity;

$$w(\mathbf{x}, \mathbf{x}') = N(\mathbf{x}; \|\mathbf{x} - \mathbf{x}'\|, \sigma_x) N(\mathbf{x}; \|\mathbf{u}(\mathbf{x}) - \mathbf{u}(\mathbf{x}')\|, \sigma_m). \tag{14}$$

Here we use $\sigma_x = 6.9$, $\sigma_m = 3.6$ and restrict $\mathbf{x}'$ to lie within 18 pixels from $\mathbf{x}'$ but within the same region r_f, so as to only smooth the flow inside the same "object". The filtered optic flow $\bar{\mathbf{u}}$ serves to predict the mask m'_{r_f} at frame $f + 1$ of region r_f. Hence the similarity between the predicted and the masks of the N regions at frame $f+1$, $\left\{ m_{r^j_{f+1}} \right\}_{j=1}^{N}$, provides the cost of linking the regions:

$$s_{r^i_f - r^j_{f+1}} = \frac{m'_{r^i_f} \cdot m_{r^j_{f+1}}}{m'_{r^i_f} + m_{r^j_{f+1}}} \tag{15}$$

The hierarchical segmentation of [1] defines semantic regions over 255 coarse to fine levels. Dynamic programming is employed to find all the possible region paths, over all frames and all levels. In particular, forward/backward dynamic programming assures that a region has unique past-to-future links. This procedure results in multiple trajectories of dense regions at different resolutions, e.g. the shirt, torso and silhouette of a pedestrian.