MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Probabilistic Motion Segmentation of Videos for Temporal Super Resolution

Arasanathan Thayananthan * Masahiro Iwasaki † R. Cipolla *

* University of Cambridge
Department of Engineering
Cambridge, CB2 1PZ, UK
{at315|cipolla}@eng.cam.ac.uk

† Panasonic Europe Ltd.
18a Sheraton House, Castle Park
Cambridge, CB3 0AX, UK
Masahiro.Iwasaki@eu.panasonic.com

Abstract

A novel scheme is proposed for achieving motion segmentation in low-frame rate videos, with application to temporal super resolution. Probabilistic generative models are commonly used to perform unsupervised motion segmentation in videos. While they provide a general and elegant framework, they are hampered by severe local minima problems and often converge to inaccurate solutions, when there are more than one foreground object in videos. This paper proposes a scheme, where ***discriminative*** *global constraints are enforced in combination with generative learning, to overcome the local minima problems. We demonstrate the effectiveness of the proposed scheme by learning the appearances and motions of multiple objects from a low frame rate video with a small number of frames.*

1. Introduction

This paper describes a scheme for achieving *temporal super resolution* in monocular videos. The term temporal super resolution in this context implies the synthesis of a number of intermediate frames such that the frame rate of the video is increased, i.e. from 5Hz video to 30Hz. It is impossible to achieve high spatial and temporal resolutions simultaneously in videos due to the physical limitations of cameras. Recently there have been a considerable research interest [1, 4, 5] in obtaining videos with high-temporal-resolution algorithmically, from low-frame-rate, high-spatial-resolution videos.

One way to achieve temporal super resolution is to identify moving objects in the low-frame-rate video, learn their motions, shapes and appearances, as well as occlusions. Once these parameters are learned, they can be used to render the intermediate frames. Learning needs to be done in an unsupervised manner, as we do not have prior knowledge about the contents of the video. Discriminative learning techniques usually need training data and cannot be used in the conventional way for our purpose. Generative models, on the other hand, explain how a video is constructed given the constituent parts of a video such as background, foreground objects and their motion models. Unsupervised learning in generative models perform the inverse process i.e. learning the constituents parts from the video data. Figure 1(a) illustrates this idea. It is usually assumed that the number of foreground objects and their motions types are known apriori, to construct the generative model.

A commonly used generative model of videos is the layer-based representation [6], where the 3D scene is decomposed into a number of 2D segments in layers. Recently Pawan Kumar et al. [3] have shown that they are able to learn and segment simple articulated motion such as fronto-parallel walking using the layer-based representation. The problem is formulated as that of labeling each pixel in the video to one of the rigidly moving objects. First they approximate the appearances and the shapes of objects by grouping pixels which have moved rigidly from frame to frame. A multi-way graph-cut is used to refine the appearances further. However, the initial grouping of pixels depend on a large number of manually set parameters, non-reversible hard decisions and a sequence of ad-hoc clustering techniques. While they have produced impressive results on a number of video sequences, it will be difficult to generalize the method, as they do not take into consideration the uncertainty in the intermediate decisions of the algorithm.

Jojic and Frey [2] introduced Bayesian probabilistic framework for the layer-based generative model. In this setup both hidden variables and model parameters are assigned probability distributions to capture the uncertainty in a principled manner. They used this framework to segment fronto-parallel translation in videos. Winn and Blake [7] extended this framework to segment a single affine motion in front of a static background in video sequences. Both methods formulate the problem as that of learning an approximate posterior distribution of the unknown variables and use variational Bayes [8] to learn it.

Probabilistic methods are attractive in an unsupervised learning scenario, since they estimate all model parameters from the data, taking into consideration the uncertainty associated with the model. However, they often assume simple likelihood models to maintain the computational tractability of the problem. These simple likelihoods models may not be enough to capture the complexity of the video data in some problems and leads to local minima solutions. Our implementations of Winn and Blake's method[7] often failed to converge, when there are more than one foreground object in the video. In this paper, we propose a scheme where additional image constraints are introduced within the generative model to reduce the chance of the solution falling into a local minima. The parameters of this additional constraints are learned discriminatively in contrast to the learning of the other variables in the model.

2 Proposed Framework

This section details the existing [6, 2, 7]layered generative model of the videos and our proposed extension. The method is motivated using a video example from [2] where two foreground objects are undergoing image plane translation in front of a stationary background. However, the framework is general and can be extended to handle videos with any number of foreground objects undergoing more complex motions.

2.1 Generative model of the video

The figure 1(a) illustrates the generative model for a video with two foreground objects. The variables inside the rectangle are repeated for each of the N frames in the video. The canonical foreground appearances $\mathbf{f}_1$, $\mathbf{f}_2$, background appearance $\mathbf{b}$, the mask priors $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are common to all the frames. Translations $\mathbf{T}_1$ and $\mathbf{T}_2$ moves the canonical appearances of the foreground objects and the mask priors to the correct locations in the frame. Mask $\mathbf{m}$ indicates the object the pixel belongs to, and has a prior given by transformed $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$. The difference between the generated image and the observed image $\mathbf{x}$ is modeled by the noise variable β. Let $\boldsymbol{\phi} = \{\mathbf{f}_1, \mathbf{f}_2, \boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \mathbf{b}, \mathbf{T}_1, \mathbf{T}_2, \beta\}$. Posterior distribution is formulated as

$$p(\mathbf{m}, \boldsymbol{\phi}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{m}, \boldsymbol{\phi})p(\mathbf{m}|\boldsymbol{\phi})p(\boldsymbol{\phi}), \tag{1}$$

where the likelihood is modeled as

$$p(x|m, \boldsymbol{\phi}) = \begin{cases} \mathcal{N}(\mathbf{x}|\mathbf{b}, \beta) & \text{if} \quad m = 0 \\ \mathcal{N}(\mathbf{x}|\mathbf{T}_1\mathbf{f}_1, \beta) & \text{if} \quad m = 1 \\ \mathcal{N}(\mathbf{x}|\mathbf{T}_2\mathbf{f}_2, \beta) & \text{if} \quad m = 2 \end{cases} \tag{2}$$

Here $m = 0, 1, 2$ indicates the pixel belonging to the background, foreground object 1 and foreground object 2, respectively. Occlusion of the pixels is modeled as

$$p(m|\boldsymbol{\phi}) = \begin{cases} (1 - \mathbf{T}_1\boldsymbol{\pi}_1)(1 - \mathbf{T}_2\boldsymbol{\pi}_2) & \text{if} \quad m = 0 \\ (1 - \mathbf{T}_2\boldsymbol{\pi}_2)(\mathbf{T}_1\boldsymbol{\pi}_1) & \text{if} \quad m = 1 \\ (\mathbf{T}_2\boldsymbol{\pi}_2) & \text{if} \quad m = 2 \end{cases} \tag{3}$$

which simply describes the fact that the object 2(the foremost layer to the camera) is always visible and the object 1 can be occluded by object 2 and the background can be occluded by both object 1 and object 2. The non-informative prior distribution over the unknown variables is given by $p(\boldsymbol{\phi})$. Exact inference of $p(\mathbf{m}, \boldsymbol{\phi}|\mathbf{x})$ is intractable and it needs to be learned using approximate methods. Stochastic techniques such as MCMC methods are commonly used to find an non-parametric approximation of the posterior distribution (e.g. particle filters). These methods rely on various sampling techniques and have slow convergence rates in high-dimensional search spaces. In this paper we use variational methods [8], which approximate $p(\mathbf{m}, \boldsymbol{\phi}|\mathbf{x})$ by a tractable parametric distribution $Q(\mathbf{m}, \boldsymbol{\phi})$ and is deterministic in nature. We use the following factorized distribution as the approximation for $p(\mathbf{m}, \boldsymbol{\phi}|\mathbf{x})$.

$$\begin{aligned} Q(\mathbf{m}, \phi) &= Q(\mathbf{m})Q(\boldsymbol{\phi}) \\ &= Q(\mathbf{m})Q(\mathbf{f}_1)Q(\mathbf{f}_2)Q(\mathbf{b})Q(\mathbf{T}_1)Q(\mathbf{T}_2) \\ &\quad Q(\boldsymbol{\pi}_1)Q(\boldsymbol{\pi}_2)Q(\beta) \end{aligned} \tag{4}$$

Here $Q(\mathbf{f}_1)$,$Q(\mathbf{f}_2)$,$Q(\mathbf{b})$ are gaussian, $Q(\mathbf{m})$,$Q(\mathbf{T}_1)$,$Q(\mathbf{T}_2)$ are discrete, $Q(\boldsymbol{\pi}_1)$,$Q(\boldsymbol{\pi}_2)$ are beta distributions and $Q(\beta)$ is a gamma distribution. A optimal set of parameters of $Q(\mathbf{m}, \boldsymbol{\phi})$ is obtained by minimizing the Kullback-Leibler (KL) divergence between $Q(\mathbf{m}, \boldsymbol{\phi})$ and $p(\mathbf{m}, \boldsymbol{\phi}|\mathbf{x})$. The KL-Divergence between these two distributions is defined as,

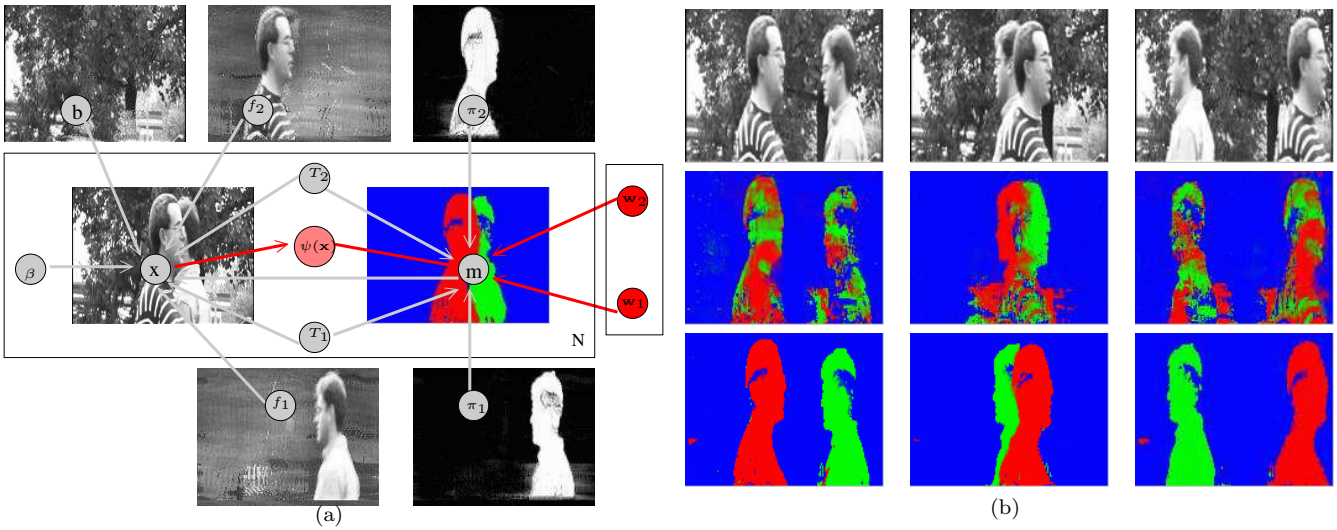

Figure 1: **(a) Generative model and Discriminative constraints.** *The figure illustrates the generative model for a video with two foreground objects. The variables inside the rectangle are repeated for each of the N frames in the video. The canonical foreground appearances* $\mathbf{f}_1$, $\mathbf{f}_2$, *background appearance* $\mathbf{b}$,*and the mask priors* $\boldsymbol{\pi}_1$ *and* $\boldsymbol{\pi}_2$ *are common to all the frames. Translations* $\mathbf{T}_1$ *and* $\mathbf{T}_2$ *moves the canonical appearances of the foreground objects to the correct locations in the frame. Mask* $\mathbf{m}$ *indicates the object the pixel belongs to, and has a prior given by transformed* $\boldsymbol{\pi}_1$ *and* $\boldsymbol{\pi}_2$. *The difference between the generated image and the observed image* $\mathbf{x}$ *is modeled by the noise variable* β. *The proposed new nodes and edges are shown in red. The feature vector* $\boldsymbol{\Psi}(\mathbf{x})$ *provides a representation of color at every image pixel. The weights* $\mathbf{w}_1$ *and* $\mathbf{w}_2$ *model the global color constraints of object 1 and object 2.* **(b) Learning results: with and without global color constraints** *The first row shows three frames from the video. Corresponding masks in the second row are learned using only the generative model without the proposed global color constraints. The third row shows the mask variables learned with both the generative model and the global color constraints.*

$$\sum_{\mathbf{m}} \int_{\boldsymbol{\phi}} Q(m)Q(\boldsymbol{\phi}) \log(p(\mathbf{m}, \boldsymbol{\phi}|\mathbf{x}))d\boldsymbol{\phi} - \sum_{\mathbf{m}} Q(\mathbf{m}) \log(Q(\mathbf{m})) - \int_{\boldsymbol{\phi}} Q(\boldsymbol{\phi}) \log(Q(\boldsymbol{\phi}))d\boldsymbol{\phi} \tag{5}$$

The above expression is minimized sequentially with respect to each of the parameters of $Q(\mathbf{m}, \boldsymbol{\phi})$ and the procedure is iterated until convergence. Further details of variational approximation scheme are omitted due to lack of space. Interested readers are referred to [8].

2.2 Discriminative Global Constraints

Our experiments showed that the variational approximation, as described in the previous section, did not converge to the correct solution due to local minima problems (figure 1(b),row 2). The simple likelihood model (equation 2) is not sufficient to capture the complex appearances of multiple objects in the video. We introduced additional constraints to the above model by setting up a discriminative prior on the mask variable $\mathbf{m}$ based on the RGB color variation of pixels, in the following form.

$$p(\mathbf{m}|\boldsymbol{\Psi}(\mathbf{x}), \mathbf{w}_1, \mathbf{w}_2) = \begin{cases} \frac{1}{Z_c} & \text{if} \quad m = 0 \\ \frac{1}{Z_c} \exp(-\mathbf{w}_1 \boldsymbol{\Psi}(\mathbf{x})) & \text{if} \quad m = 1 \\ \frac{1}{Z_c} \exp(-\mathbf{w}_2 \boldsymbol{\Psi}(\mathbf{x})) & \text{if} \quad m = 2 \end{cases} \tag{6}$$

where,

$$Z_c = 1 + \exp(-\mathbf{w}_1 \boldsymbol{\Psi}(\mathbf{x})) + \exp(-\mathbf{w}_2 \boldsymbol{\Psi}(\mathbf{x})). \tag{7}$$

The feature vector $\boldsymbol{\Psi}(\mathbf{x})$ is obtained at every pixel and consists of kernelized distances from the pixel's RGB value to a number of centers in the RGB color space. The centers are obtained by clustering RGB values of a number of sample pixels from the input frames. For the video example in this paper, we used 50 such centers to represent the color variation. The weights $\mathbf{w}_1$ and $\mathbf{w}_2$ model the distribution of the color

in the first and second objects in the video. These news constraints are shown in red color in the graphical model in figure 1(a). The modified overall posterior distribution is now given by

$$p(\mathbf{m}, \boldsymbol{\phi}, \mathbf{w}_1, \mathbf{w}_2 | \mathbf{x}, \boldsymbol{\Psi}(\mathbf{x})) \propto p(\mathbf{x}|\mathbf{m}, \boldsymbol{\phi}) p(\mathbf{m}|\boldsymbol{\phi}) p(\boldsymbol{\phi}) \\ p(\mathbf{m}|\boldsymbol{\Psi}(\mathbf{x}), \mathbf{w}_1, \mathbf{w}_2) p(\mathbf{w}_1) p(\mathbf{w}_2) \tag{8}$$

where, the prior distributions $p(\mathbf{w}_1)$ and $p(\mathbf{w}_2)$ can be either uniform or broad gaussian distributions. Using the KL divergence minimization outlined above it can be shown that at each variational iteration, optimum values of $\mathbf{w}_1$ and $\mathbf{w}_2$ can be obtained by

$$\frac{\partial}{\partial \mathbf{w}_1} \Upsilon(\mathbf{w}_1, \mathbf{w}_2) = 0 \qquad \frac{\partial}{\partial \mathbf{w}_2} \Upsilon(\mathbf{w}_1, \mathbf{w}_2) = 0 \tag{9}$$

where

$$\Upsilon(\mathbf{w}_1, \mathbf{w}_2) = \sum_{\mathbf{m}} Q(\mathbf{m}) log \left\{ p(\mathbf{m}|\boldsymbol{\Psi}(\mathbf{x}), \mathbf{w}_1, \mathbf{w}_2) p(\mathbf{w}_1) p(\mathbf{w}_2) \right\} \tag{10}$$

Equations(9,10) and the form of $p(\mathbf{m}|\boldsymbol{\Psi}(\mathbf{x}), \mathbf{w}_1, \mathbf{w}_2)$, as described in equation (6), leads to logistic regression learning of $\mathbf{w}_1$ and $\mathbf{w}_2$, with the probabilities $Q(\mathbf{m})$ acting as a training signal. The weights parameters are learned using information from all the frames and can prevent the mask variable falling into local minima in individual frames as shown in figure 1. Note that the learning and enforcement of the discriminative constraints are done in a completely unsupervised manner and fits well within the variational framework as shown by the above derivation. The method is not sensitive to the feature vector $\boldsymbol{\Psi}(\mathbf{x})$. Different feature vectors that can capture color variability can be also used with equal success. Figure 1(b) and figure 2 illustrates the advantages of introducing the discriminative global color constraints to the standard generative model.

2.3 Frame Synthesis and Temporal Super resolution

Once the canonical appearances, shapes and motions for each frame are learned, intermediate frames are synthesized by first interpolating the motion between frames using a motion model. Let T_1^* and T_2^* be the interpolated motions for an intermediate frame, whose pixels values $\mathbf{x}^*$ are then obtained by

$$\mathbf{x}^* = (1 - \mathbf{T}_1^* < \boldsymbol{\pi} 1 >)(\mathbf{T}_2^* < \boldsymbol{\pi}_2 >)(< \mathbf{b} >) \\ + (1 - \mathbf{T}_2^* < \boldsymbol{\pi}_2 >)(\mathbf{T}_1^* < \boldsymbol{\pi}_1 >)(\mathbf{T}_1^* < \mathbf{f}_1 >) \\ + (\mathbf{T}_2^* < \boldsymbol{\pi}_2 >)(\mathbf{T}_2^* < \mathbf{f}_2 >) \tag{11}$$

Here operator $< . >$ is the expected value of the variable. For example, the expected value of the first object's appearance is given by $< \mathbf{f}_1 >= \int_{\mathbf{f}_1} \mathbf{f}_1 Q(\mathbf{f}_1) \mathbf{df}_1$.

3 Experiments

The method was tested on 40 frames of a real 15 Hz video (resolution 160x120), where two people cross in front of a stationary background. Variational optimization using the generative model alone is not sufficient to learn the appearances and motions of the objects in the video as shown in figure 2(b) second row, even when all 40 frames were used. Including the discriminative global color constraints on the mask variable improves the quality of the learning, figure 2(b) third row. The learned canonical appearances shown in figure 1(a) are obtained with global color constraints. We also subsampled the original video and created lower frame rate videos (3.75 Hz, 11 frames). We then learned the model parameters (with global color constrains) and used them to synthesize intermediate frames in order to achieve temporal super resolution. It can be seen from figure 2(c) that introducing the color constraints allows us to avoid the local minima even with a small number of input frames.

4 Conclusion

This paper introduced a scheme where discriminative global constraints were used to avoid local minima solutions during unsupervised learning of generative models of videos. The scheme was demonstrated by performing four times higher temporal super resolution from a low-frame-rate video with a small number of frames. It should be noted that the scheme is not limited to color constraints only. We can also easily introduce other types of constraints (such as smoothness, object locality and edges) within this framework. In future we intend to demonstrate the system in video with more complex transformation such as affine using additional discriminative constraints.

Acknowledgments This work was supported by Panasonic Europe Ltd.

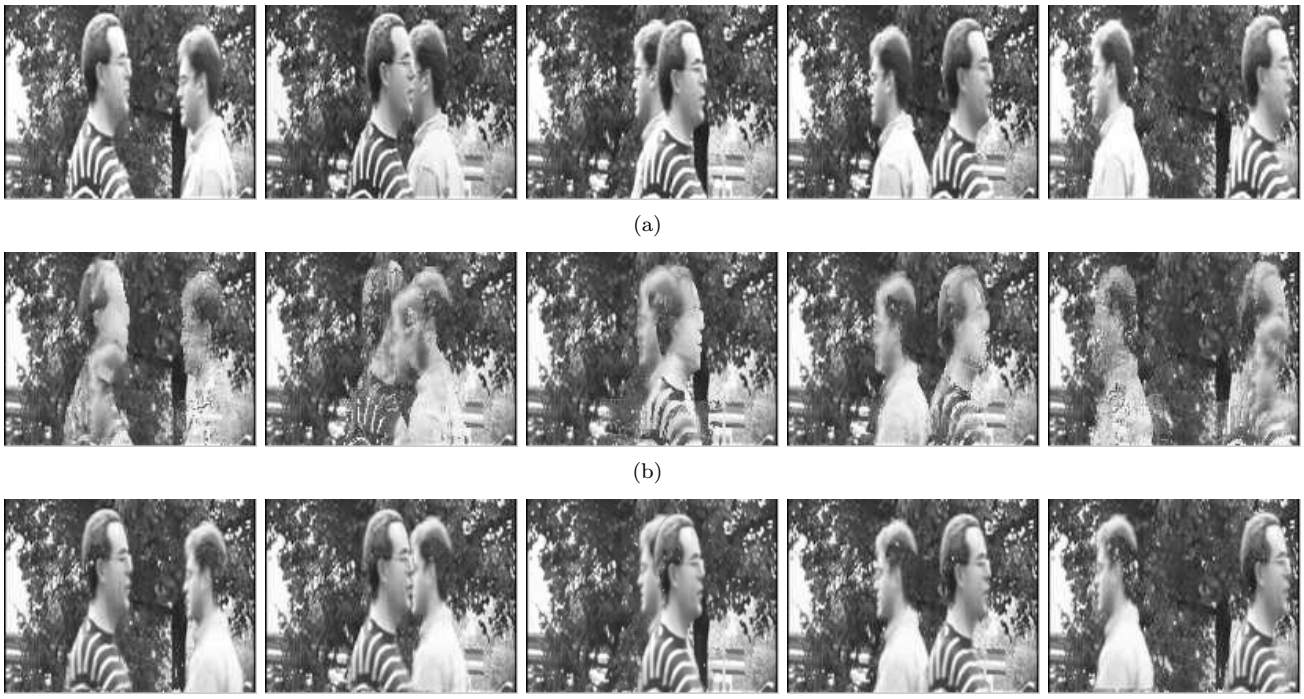

Figure 2: **Frame synthesis and Temporal Super Resolution. (a)** *shows real frames from a 40 frame video sequence.* **(b)** *Synthesized frames created after learning the appearances and the motions of the objects in the video using the generative model shown in figure 1 (a), but without enforcing the global color constraints. The full video (with 40 frames) was used for learning.* **(c)** *Synthesized frames created after learning the appearances and the motions using the generative model with global color constraints. A sub-sampled version of the full video (with only 11 frames) was used for learning. From these 11 frames we were able to synthesize 30 intermediate frames giving a temporal super resolution from 3.75Hz to 15Hz.*

References

[1] V. Cheung, B. Frey, and N. Jojic. Video Epitomes. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2005.

[2] N. Jojic, and B. Frey. Learning flexible sprites in video layers. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2001.

[3] M. Pawan Kumar, P. H. S. Torr, and A. Zisserman. Learning layered motion segmentations of video. In *Proc. 10th Int. Conf. on Computer Vision*, 2005.

[4] Y. Wexler, E. Shechtman, and M. Irani. Space-time video completion. In *Proc. Conf. Computer Vision and Pattern Recognition*, 2004.

[5] M. Wilczkowiak, G.J. Brostow, B. Tordoff, and R. Cipolla. Hole filling through photomontage. In *Proc. British Machine Vision Conference*, 2005.

[6] J. Y. A. Wang, and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*,vol. 3, no. 5, pp. 625-638, 2004.

[7] J. Winn, and A. Blake. Generative affine localisation and tracking. In *NIPS*, 2004.

[8] J. Winn, and C.M. Bishop. Variational Message Passing. *Journal of Machine Learning Research*,vol. 6, pp. 661-694, 2005.