A Multiple View Layered Representation for Dynamic Novel View Synthesis

Keith Connor and Ian Reid
Department of Engineering Science
University of Oxford
Oxford, OX1 3PJ, UK
[krc,ian]@robots.ox.ac.uk

Abstract

We propose a multiple view layered representation for tracking and segmentation of multiple objects in a scene. Existing layered approaches are dominated by the single view case and generally exploit only motion cues. We extend this to integrate static, dynamic and structural cues over a pair of views. The goal is to update coherent correspondence information sequentially, producing a multi-object tracker as a natural byproduct. We formulate a MAP solution for estimating layer parameters which are consistent across views, with the EM algorithm used to determine both the hidden segmentation labelling and motion parameters. A persistent representation of occupancy is maintained in spite of occlusion without enforcing a particular parametric shape model. An immediate application is dynamic novel view synthesis, for which our layered approach offers a direct and convenient representation.

1 Introduction

The layered representation has grown from a need to represent and describe natural scenes compactly. A video sequence, for example, may be represented in terms of a small number of regions and their associated motions [11]. Layers have mainly found use in the representation of monocular video sequences, typically for applications concerned with video coding.

A new layered representation suitable for multiple view descriptions of dynamic scenes is presented. This is motivated by the Novel View Synthesis (NVS) problem which requires correspondence information across views. Our approach seeks only implicit 3D information, i.e. correspondence, and is therefore an image-based approach not requiring calibration. We introduce the concept of layer aspects in order to model view dependent layer properties and augment this with a global visibility representation for each view which handles inter-layer occlusions. Layer parameters are computed sequentially for a pair of views conditioned on the layer parameters at the previous time. This amounts to tracking multiple objects in two views where, as will be shown, the layer parameters can be used to render consistent novel views over time, i.e. Dynamic Novel View Synthesis.

BMVC 2003 doi:10.5244/C.17.72

In this paper we apply the method to view interpolation, but it is equally applicable to more general synthesis methods [4].

1.1 Related Work

The typical conception of a layer is based on translating object-like properties to the image plane. That is, objects tend to have spatial continuity and coherent motion, and thus a layer should have the same properties. Difficulties are encountered because projection discards depth information and complications such as occlusion result.

The most common forms of layered model encountered in the literature are designed for the single view case. The computation of layers has been formulated as estimating the parameters of a number of global motion models and their regions of support. A region of support is simply a subset of image pixels which *belong* to a layer.

Wang and Adelson [10], robustly compute affine motion parameters over an arbitrary grid of patches and proceed to cluster motion and re-evaluate both the number and extent of the layers. More recently, approaches by [5] [1], employ a probabilistic mixture model formulation to compute the maximum likelihood layer parameters and the number of layers. The EM-algorithm [6] has become the standard tool for estimating layer parameters because it enables the simultaneous computation of segmentation and motion.

A particular variant among previous approaches is whether or not occlusion is fully accounted for. The persistent representation of a layer's occupancy in spite of occlusion is key for tracking and is exploited by Jepson *et al.* [7], where a strong shape model is employed. Tao *et al.* model a layer's shape by a Gaussian spatial prior but this serves more as a segmentation (i.e. visibility) prior rather than an occupancy prior and thus does not explicitly consider occlusion.

Most previous approaches compute motion layers for a single view of a dynamic scene, while other less prolific work considers structural layers. The work of [2] and [9] consider two-views of a scene in order to extract 3D layers, where the transformations between views is due to structure rather that dynamic object motion. In contrast, our work considers both motion and structure.

2 A Layered Representation

This section introduces the chosen layered representation and describes the concepts governing a generative model. We describe a layer in terms of spatial extent, appearance, alignment and visibility. The latter is particularly important for an image-based representation. These terms will be defined and quantified in the rest of this section.

The layered model is composed of $N+1$ layers, i.e. the background layer and N foreground layers. In general each layer will have a number of aspects where an aspect is defined to be a layer's representation in a particular view. Figure 1 illustrates the general layered representation for multiple layers with multiple aspects.

In particular, the the layered model at a time t is denoted as $L_t = (O_t, A_t, \Phi_t)$, where, in order, the parameters represent: occupancy, appearance and alignment, and are independent for each layer. In contrast, visibility is a direct function of all layers and is also view (aspect) dependent. The concept of an aspect is necessary in order to allow for view dependent visibility. Alignment parameters Φ_t^i encode the mapping (correspondence) between aspects of the i'th layer.

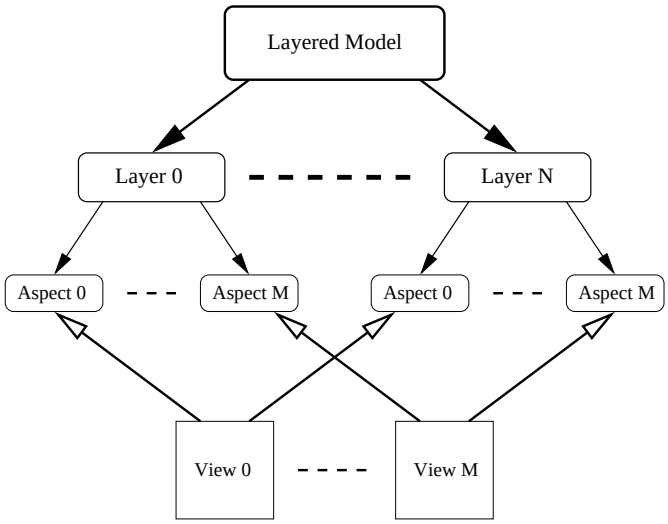

Figure 1: Multiple aspect layered model. Layers are composed of a number of aspects which correspond to the layer's representation in a particular view. In general, the model has N layers and each layer has M aspects.

2.1 Generative model

An image I obtained at time t is generated from a mixture distribution, where the value of each pixel $I_t(x)$ is sampled from the following distribution:

$$P(I_t(x)) = \sum_{i=0}^{N} P(I_t(x) \,|\, V_t(x) = i)\, P(V_t(x) = i)\,, \qquad \sum_{i=0}^{N} P(V_t(x) = i) = 1 \tag{1}$$

where, $V_t(x)$ is a visibility indicator variable which evaluates to i if the i'th layer is visible at x.

The likelihood distribution for an observed image intensity, conditioned on the visibility of a particular layer, constitutes an appearance model for that layer. In particular, the colour distribution at pixel x given that layer i is visible is precisely the appearance distribution, denoted by a random variable A_t^i, such that $P(A_t^i \,|\, x) = P(I_t(x) \,|\, V_t(x) = i)$. This may be, for example, a colour histogram, a parametric mixture model, or a complete spatially dependent statistical model. Thus, equation 1 is effectively a composition of aspect appearances.

In the following, depth ordering amongst layers is denoted implicitly through layer indices, where layer 0 is taken to be the background layer: a special layer which is assumed to occupy every pixel in both views. Whereas visibility is mutually exclusive at pixel sites, occupancy is not, and in fact visibility is a direct function of the occupancies. Further, knowing the depth relationship among layers:

$$P(V_t(x) = i) = P(O_t^i(x)) \prod_{j=i+1}^{N} (1 - P(O_t^j(x))) \tag{2}$$

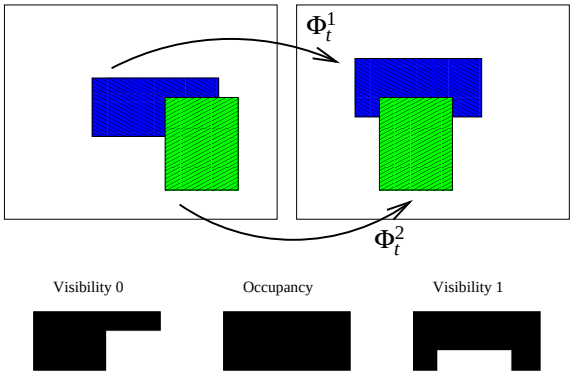

Figure 2: A layer is represented by persistent occupancy and view-dependent visibility probability densities. A layer is partially occluded in two views (left) and the corresponding representation is shown (right).

This simply expresses that for a layer to be visible at a pixel x, it must occupy x, and layers that could occlude it, must not occupy x. The foremost layer visibility probability is therefore equal to its occupancy probability, $P(V_t(x) = N) = P(O_t^N(x))$, and the background layer visibility is dependent only on the other layer occupancies.

In this paper we concentrate on the situation for two views of a dynamic scene and therefore each layer has two aspects. Figure 2 illustrates the how occlusion is modelled in and across views by maintaining a persistent occupancy representation.

3 Computing layer parameters

We wish to compute the parameters of the layered model that maximise the posterior likelihood of the current image pair given the previous layer parameters. Bayes rule gives:

$$P(L_t \,|\, I_t, L_{t-1}) \propto P(I_t \,|\, L_t, L_{t-1})\, P(L_t \,|\, L_{t-1}) \tag{3}$$

Maximising this posterior with respect to the layer parameters is difficult in general, however this fits well into the EM-algorithm [6]. We assume the hidden data are the visibilities over the current images, i.e. variables that indicate which layer is responsible for generating each pixel (see section 2.1). Thus, the E-step corresponds to obtaining the visibility posteriors given the current estimate of the layer parameters. The M-step consists of maximising the following objective function over the layer parameters.

$$Q(L_t, L_t') = E\left[\ln P(I_t, V \,|\, L_t, L_{t-1}) \,|\, I_t, L_t', L_{t-1}\right] + \ln P(L_t \,|\, L_{t-1}) \tag{4}$$

The following sections describe the individual steps taken to optimise this objective function which constitute a Generalised EM-algorithm. Figure 3 gives an overview of the steps taken for each iteration.

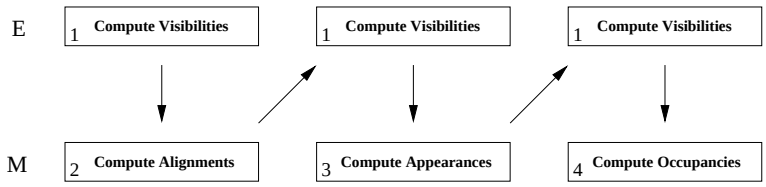

Figure 3: Overview of steps in computation.

3.1 Temporal update for layer priors

Given the layer parameters at the previous time step, L_{t-1}, we compute a prior on the parameters a time t. In order to do this, the parameters of the chosen motion model are computed between successive frames in each view. In our case, we compute an affine motion for both aspects of each layer, W, by minimising the following error function over the six parameters of W separately in each view:

$$\mathscr{E} = \sum_{x \in I_{t-1}} P(V_{t-1}(x) = i) \left[I_{t-1}(x) - I_t(W^i(x)) \right]^2 \tag{5}$$

This is minimised by a Gauss-Newton procedure using a weighted version of the inverse compositional algorithm proposed by [3]. Thus, we obtain estimates of the temporal motions, W_1^i and W_2^i, for each layers aspects.

In the following we drop layer indices on the understanding that we are considering a particular layer. Occupancy priors are computed by transforming the previous posteriors on occupancy via the temporal motions. This effectively shifts the occupancies into probable locations. More precisely, an occupancy prior is computed by:

$$P(O_t(x)) = \int_{x' \in I_{t-1}} P(O_{t-1}(x')) P(x') dx' \tag{6}$$

where, $E[x'] = W^{-1}x$, and $P(x')$ is the probability distribution for the location x' associated with the transformation's covariance. If the actual motion is correctly modelled by the chosen motion model and computed with complete accuracy, then $P(x' = W^{-1}x) = 1$ and the occupancies are simply transferred, in general however, $P(x')$ account for the uncertainty in the estimated motion.

The transitive relationship between a layer's motion in each view and alignment across views provides a prior estimate of the current alignment parameters for each layer:

$$\Phi_t = W_2 \Phi_{t-1} W_1^{-1} \tag{7}$$

3.2 Computing visibility

This section describes step 1 from figure 3. The posterior visibilities are computed for each view by computing the expectation given the current estimates of the layer parameters. Bayes' rule gives:

$$P(V_t(x) = i \,|\, I_t, L_t, L_{t-1}) = \frac{P(I_t \,|\, V_t(x) = i, L_t, L_{t-1}) \, P(V_t(x) = i \,|\, L_t, L_{t-1})}{Z} \tag{8}$$

and following from equation 2, we have, for the first view:

$$P(V_t(x) = i \,|\, I_t, L_t, L_{t-1}) = \frac{P(A_t^i)\, P(O_t^i(x)) \prod_{j=i+1}^{N} [1 - P(O_t^j(x))]}{Z} \tag{9}$$

and for the second view:

$$P(V_t'(x) = i \,|\, I_t', L_t, L_{t-1}) = \frac{P(A_t^i)\, P(O_t^i(\Phi_t^i(x))) \prod_{j=i+1}^{N} [1 - P(O_t^j(\Phi_t^j(x))]}{Z} \tag{10}$$

We also evaluate the joint distribution for visibilities which are required by step 4 in figure 3. We consider the joint distribution under the i'th alignment:

$$P(V(x) = j, V'(\Phi^i(x)) = k \,|\, I_t, I_t', L_t, L_{t-1}) = \tag{11}$$

$$P(A_t^j, A_t^k)\, P(V'(\Phi^i(x)) = k \,|\, V(x) = j)\, P(V(x) = j) \,/\, Z \tag{12}$$

where the conditional probability can take values as follows:

$$= \quad 0 \quad : \quad (j = i, k < i)\,,\ (j < i, k = i) \tag{13}$$

$$= \quad \prod_{h=i+1}^{N} [1 - P(O^h(\Phi^i(x)))] \quad : \quad (j = i, k = i) \tag{14}$$

$$= \quad P(V'(\Phi^i(x)) = j) \quad : \quad \textit{otherwise} \tag{15}$$

The choice of a suitable appearance model is to a large extent application dependent. In the present work, we represent the background appearance by a Gaussian distributions for each pixel and each foreground layer appearance by a single Gaussian distribution. The parameters are updated using a weighted sum of the current parameters and the new observations according to the visibility posteriors. Such a perversely simple model is flexible enough to cope with a useful range of interesting scenarios. In particular, the sequences of football action we present in section 5 are handled adequately. Other possible models are currently being evaluated.

3.3 Computing alignment

Each layer has affine alignment parameters, Φ_t^i, which put the layer's aspects into correspondence. Here we consider step 2 from figure 3. We compute the alignment of aspects by minimising the following error function:

$$\mathscr{E} = \sum_{x \in I_t} P(V_t(x) = i, V_t'(\phi_t^i(x)) = i) \left[I_t(x) - I_t'(\Phi_t^i(x)) \right]^2 \tag{16}$$

where, ϕ_t^i represents the current alignment and Φ_t^i represents the updated alignment. This weighting by the visibility probabilities ensures that only valid measurements are used and is therefore robust to occlusion.

3.4 Computing occupancy

Occupancy and visibility are inextricably coupled through the composition relationship of equation 2. At this point it is necessary to infer the occupancies given the expected visibilities in each view. The occupancy of layer i is obtained by marginalising the joint

distribution $P(O^i, V, V')$. We omit the extraneous conditional parameters and assume we are considering pixel x in view one and the corresponding pixel under the relevant layer alignment, $\Phi^i(x)$, in view two. It can be shown that the required marginal density can be obtained as follows:

$$P(O^i) = \sum_{j=0}^{N}\sum_{k=0}^{N} P(O^i, V=j, V'=k) \tag{17}$$

$$= \sum_{j=i}^{N}\sum_{k=i}^{N} P(O^i \,|\, V=j, V'=k)\, P(V=j, V'=k) \tag{18}$$

$$= P(V=i, V'=i) + \sum_{j=i+1}^{N} P(V=j, V'=i) + \sum_{k=i+1}^{N} P(V=i, V'=k)$$
$$+ \sum_{j=i+1}^{N}\sum_{k=i+1}^{N} P(O^i \,|\, V=j, V'=k)\, P(V=j, V'=k) \tag{19}$$

Given that a layer is occluded in both views, we can interpret the conditional probability, $P(O^i|V=j,V'=k) \;:\; j>i\,,\, k>i$, as the prior occupancy probability.

To illustrate the preceding occupancy probability: the first term accounts for layers visible in both views, the second and third terms account for layers visible only in one view and therefore occluded in the other and, the last term accounts for layers occluded in both views. The joint distribution over visibilities in two views was given in section 3.2.

4 View Synthesis

The layered representation described provides a compact and convenient rendering model for novel view synthesis. This section describes how it is exploited for the case of view interpolation. View interpolation can be formulated as: the linear interpolation of corresponding position and colour. Thus a correspondence, $x \leftrightarrow x'$, interpolated by parameter, λ, corresponds to a position, $x'' = (1-\lambda)x + \lambda x'$, with colour evaluated as:

$$I''(x'') = (1-\lambda)I(x) + \lambda I'(x') \tag{20}$$

The first step is to evaluate visibilities for the novel view. This is exactly the same process as used for the real views except that now the layer alignment parameters are interpolated to produce novel aspects. The following achieves an affine transformation which corresponds to interpolating positions $x \leftrightarrow \Phi^i(x)$:

$$\Phi^i(\lambda) = (1-\lambda)I + \lambda\Phi^i \tag{21}$$

For each layer, novel aspects are produced by transforming the occupancy probabilities into the new view. Next, in order to render the correct layer at each pixel, the visibilities are computed. Finally, we take the maximum a posteriori decision on which layer is visible at each pixel and render according to the following rule:

$$I'' = \frac{(1-\lambda)\,P(V(x)=i)\,I(x) + \lambda\,P(V'(x')=i)\,I'(x')}{(1-\lambda)\,P(V(x)=i) + \lambda\,P(V'(x')=i)} \quad : \quad i = \arg\max_j P(V''(x'')=j) \tag{22}$$

which also achieves the natural requirement that we should only use the colour from visible parts of each layer.

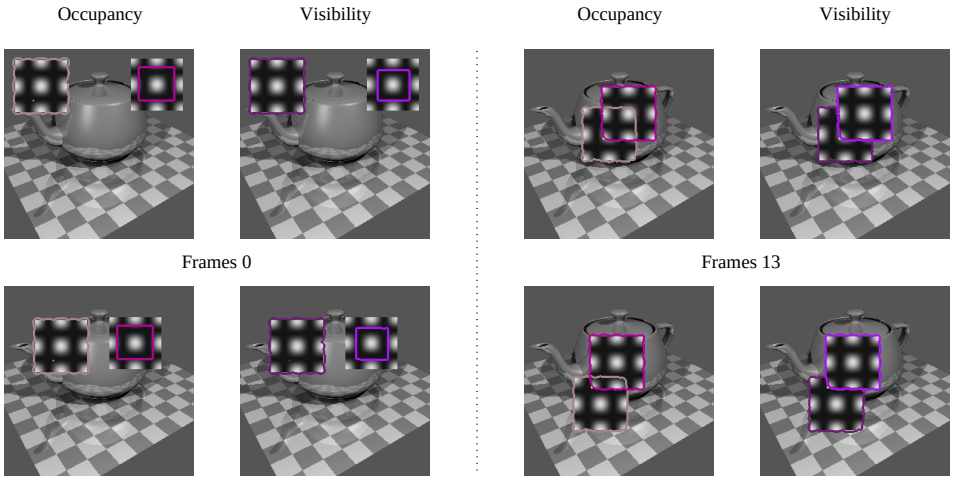

Figure 4: A pair of synthetically generated sequences with static backgrounds and two moving foreground objects. (top row) first view and (bottom row) second view. On the left is the result for the initial frame pair showing occupancy and visibility. On the right at a later time in which occlusion takes place.

5 Results

We first present an experiment using a synthetically produced pair of image sequences (figure 4). Each view is composed of a background image with two textured foreground squares which move over the background. In order to mimic the two-view situation we give the foreground elements different positions and velocities in each view. This sequence is made purposefully difficult in terms of large occlusions between foreground elements. To illustrate the states of the layered representation we draw edges where each layer's occupancy region changes from $(P(O^i) > 0.5)$ to $(P(O^i) < 0.5)$. Visibility is also illustrated by edges where the MAP determined visible label changes. This sequence is manually initialised and we deliberately underestimate the right-hand foreground object's occupancy prior in the first frame in order to show how the occupancy grows to explain the data. The large occlusion (around 50%) is handled extremely well and maintains a viable occupancy representation when not observed.

The next experiment is conducted on real data (figure 5). We obtained two video sequences of a football match, focusing on the action in the penalty area. The layers are manually initialised by the rough selection of a number of corresponding regions. Some occlusion is also apparent in this sequence as well as non-rigid motions and rapid changes in appearance. The layers are computed over one-hundred frames. There are several frames in which the goalie suffers near full occlusion, but despite the difficulties posed by this sequence, the players occupancies are well represented and tracked.

Finally, based on the computed layers from the football sequence we compute novel views in figure 6. Intermediate views at three points in time are shown. The background layer is synthesised using an alternative interpolation method [8] which reasonably models the structure. Because the cameras are fixed the structure (correspondences) for the

Figure 5: Two-view tracking of football player layers. The top and bottom rows show the first and second camera views respectively. Four frames are taken from the results over a hundred frame sequence and occupancy is illustrated by edges as explained in the text.

background can be precomputed. As can be seen, we achieve photo-realistic, dynamic novel view synthesis under occlusion and unmodelled motions.

6 Conclusion

We have introduced a new layered representation for multiple views of dynamic scenes and described an MAP solution for sequentially estimating the parameters of this model. The result is a procedure which can track multiple moving objects over a pair of views with a complete representation of salient properties. In particular, it maintains a persistent representation of occupancy in spite of occlusions and integrates measurements from each view. The motivation for this research was in obtaining a suitable model for dynamic novel view synthesis. We have shown that for this problem we can produce compelling, photo-realistic results.

In our current work we are considering several important issues. The number of layers should be obtained automatically and revised over time within the same framework by exploiting constraints from domain knowledge in association with model complexity coding concepts. Structural layers are special cases of dynamic layers in that they are static, and as such complicated backgrounds can be represented within the same layered framework. Finally, the practical extension to more than two views offers further interesting viewing constraints for computing the representation's parameters.

7 Acknowledgements

This work is supported by the European Framework 5 grant *EVENTS* [IST-1999-21125].

Figure 6: Dynamic novel view interpolation. Each row corresponds to a point in time. Each column corresponds to an interpolated view.

References

[1] S. Ayer and H. S. Sawhney. Layered representation of motion video using robust maximum likelihood estimation of mixture models and MDL encoding. In *Proc. of International Conference on Computer Vision*, pages 777–784, 1995.

[2] S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, June 1998.

[3] Simon Baker and Iain Matthews. Equivalence and efficiency of image alignment algorithms. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Dec 2001.

[4] K. Connor and I. Reid. Novel view specification and synthesis. In *Proc. 10th British Machine Vision Conference, Cardiff*, September 2002.

[5] T. Darrell and A. Pentland. Cooperative robust estimation using layers of support. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17:474–487, 1995.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc.*, 39:1–38, 1977.

[7] A. D. Jepson, D. J. Fleet, and M. J. Black. A layered motion representation with occlusion and compact spatial support. In *Proc. of European Conference on Computer Vision*, pages 692–706, 2002.

[8] Maxime Lhuillier and Long Quan. Image interpolation by joint view triangulation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Fort Collines, USA, June 1999.

[9] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated Bayesian approach to layer extraction from image sequences. In *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pages 983–990, 1999.

[10] J. Y. A. Wang and E. H. Adelson. Layered representation for motion analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 361–366, New York, June 1993.

[11] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, 3(5):625–638, Sep 1994.