Generative Model for Layers of Appearance and Deformation

Anitha Kannan[1], **Nebojsa Jojic**[2], **Brendan J. Frey**[1]
[1] Probabilistic and Statistical Inference Group, University of Toronto
[2] Microsoft Research, Redmond, WA, USA

Abstract

We are interested in learning generative models of objects that can be used in wide range of tasks such as video summarization, image segmentation and frame interpolation. Learning object-based appearance/shape models and estimating motion fields (deformation field) are highly interdependent problems. At the extreme, all motions can be represented as an excessively large set of appearance exemplars. However, a more efficient representation of a video sequence would save on frame description if it described the motion from the previous frame instead. The extreme in this direction is also problematic as there are usually causes of appearance variability other than motion. The flexible sprite model (Jojic and Frey 2001) illustrates the benefits of joint modelling of motion, shape and appearance using very simple models. The advantage of such a model is that each part of the model tries to capture some of the variability in the data until *all* the variability is decomposed and explained through either appearance, shape or transformation changes. Yet, the set of motions modelled is very limited, and the residual motion is simply captured in the variance maps of the sprites. In this paper, we develop a better balance between the transformation and appearance model by explicitly modelling arbitrary large, non-uniform motion.

1 Introduction

Our objective is to learn generative models of objects in a visual scene so that scene analysis (such as video summarization) can be efficiently performed. An important component of scene analysis involves learning object based appearance/shape models and estimate motion reliably. These two interesting problems of learning object based appearances and estimating motion are extensively studied separately even though both appearance and motion provide independent cues for estimating each other. In this work, we introduce a probabilistic generative model that unifies appearance modelling and motion estimation.

A step in this direction is reported in (Jojic and Frey 2001), as a layered extension of (Frey and Jojic 1999) for multiple objects. Here, the goal is to learn a layered density model for image formation. Given an input video sequence, the approach iteratively updates the appearances and masks of objects associated with each layer and the estimates of *global transformation(motion)* of the objects while capturing the residual motion in the variance maps of the sprites. Despite this appearance flexibility, the model requires an excessive number of appearance classes to capture many types of nonuniform large motions for which the translational variable is not a sufficient descriptor. We describe a new generative model for layered image formation that simultaneously learns deformation-invariant appearances and infer complex deformation fields. We use variational inference and generalized EM for learning and present results on flow computation, image segmentation and frame interpolation.

2 Related work

In a scene with multiple objects, approaches to motion estimation that operate on *matching* patches from one image to another (Lucas and Kanade 1981) under perform at the boundary regions due to occlusions and disocclusions. A good appearance model enables effective handling of boundary regions. On the other hand, objects can undergo complex deformations, and motion provides useful cues to learn appearances (Black and Jepson 1996). Thus, the estimation of appearances and motion should be done in tandem. A popular approach to this is to use a layered representation

in which we decompose a 3 dimensional scene into a set of 2 dimensional layers.

One such layered formalism is based on mixture models (Ayer and Sawhney 1995), (Jepson and Black 1993), (Weiss and Adelson 1996) in which each pixel is assigned probabilistically to one of several layers. When multiple objects are moving in a scene, there is a fair amount of occlusions and disocclusions, and without proper appearance models for the objects in the scene, it is extremely difficult to find the boundaries of the object.

In (Black and Fleet 2000), a framework for modelling motion discontinuities is presented. In this work, the foreground and background are separated by a straight edge within a single, fixed window in the image sequences. The image sequence within the window is modelled by a generative model that predicts the image at time t from the image at time $t-1$ using unknown state variables that describe the location of the edge and the motions of the foreground and background. An algorithm based on particle filtering is used to infer the location of the edge, motion vectors for the foreground and the background at each time step. Again, this approach does not have explicit model for appearances of the objects, but instead relies on straight edge to differentiate foreground and background pixels within a small window. Moreover, for complex object shapes, a single edge may not be sufficient to differentiate the two layers.

In (Jojic and Frey 2001), a generative model framework is used to automatically learn layers of "flexible sprites", which are probabilistic 2-dimensional appearance maps and masks of moving, occluding objects. An important assumption of this model is that pixels belonging to a sprite move with the same velocity (for instance, uniform translation). For many interesting video sequences, this assumption is too rigid.

In (Frey, Jojic and Kannan 2003), we suggest linearizing the transformation manifold locally. This approach has two drawbacks - it requires an additional global transformation for finding the position and often a linear manifold is not sufficient to capture *large* complex deformation. The use of low-frequency wavelets for smooth deformation fields (Jojic et al. 2001) suffers from the same problem.

3 Flexible sprites with deformation fields

Fig. 1 shows the hierarchical generative model that describes the process involved in two-layer image formation. The statistical generative process is as follows: For each layer, an appearance and a mask are generated from appropriate prior distributions associated with object classes. We sample deformation vectors for each pixel. The deformation field is then applied to both the appearance and the mask. The position variables are randomly selected and the appropriate latent images shifted in accordance. The final image is composed from the layers according to the masks, which can be either continuous or discrete.

At this juncture, we would like to point out that the deformation field is fully expressive and nonlinear, and the model without the position variable can still capture well the correct global motion. We have added the shift variable only to serve the purposes of regularization and speedup of computation. There are too many relatively good solutions to arbitrary matching patches in the mean image and the observation. The existence of the shift variable limits the search space for the deformation field estimation, *and* regularizes the search.

This is also the main difference from our earlier work (Frey, Jojic and Kannan 2003), where we used a linearizing manifold locally. To make this linearization work, an added nonlinearity is needed, and for that purpose, we used discrete shifts.

4 A generative model for occluded patches in motion

Although the following discussion applies to an arbitrary number of layers, we consider for simplicity a two-layer model, consisting of a foreground and a background. We treat foreground appearance (denoted by $\mathbf{f}$) and background appearance ($\mathbf{b}$) as parameters that apply to an entire sequence. (In the full model, there are several layers and several appearance classes that can occupy them).

We associate with the foreground layer a binary mask $\mathbf{m}$ of the same size as $\mathbf{f}$ such that $\mathbf{m}_i$=1 indicates that the corresponding pixel is a foreground pixel. Let the probability that $\mathbf{m}_i = 1$ be α_i so that

$$P(\mathbf{m}) = \prod_i \alpha_i^{m_i}(1-\alpha_i)^{(1-m_i)}.$$

Although the multiplicative alpha map we used in some of our previous papers is attractive for modeling mixed pixels, the binary mask tends to allow for a more robust inference (Frey and Jojic 2004)(Williams and Titsias 2003).

Each pixel in the latent images undergoes a deformation. In this paper, we use a discrete coordinate system for clarity, although techniques for sub-pixel inference and multi-scale search can be used. A priori the foreground and background motion vectors, $\mathbf{U}$ and $\mathbf{V}$ are independent and follow uniform distribution denoted by $P(\mathbf{U})$ and $P(\mathbf{V})$. We can favor smaller motions by using, for instance, a Gaussian prior on displacement.

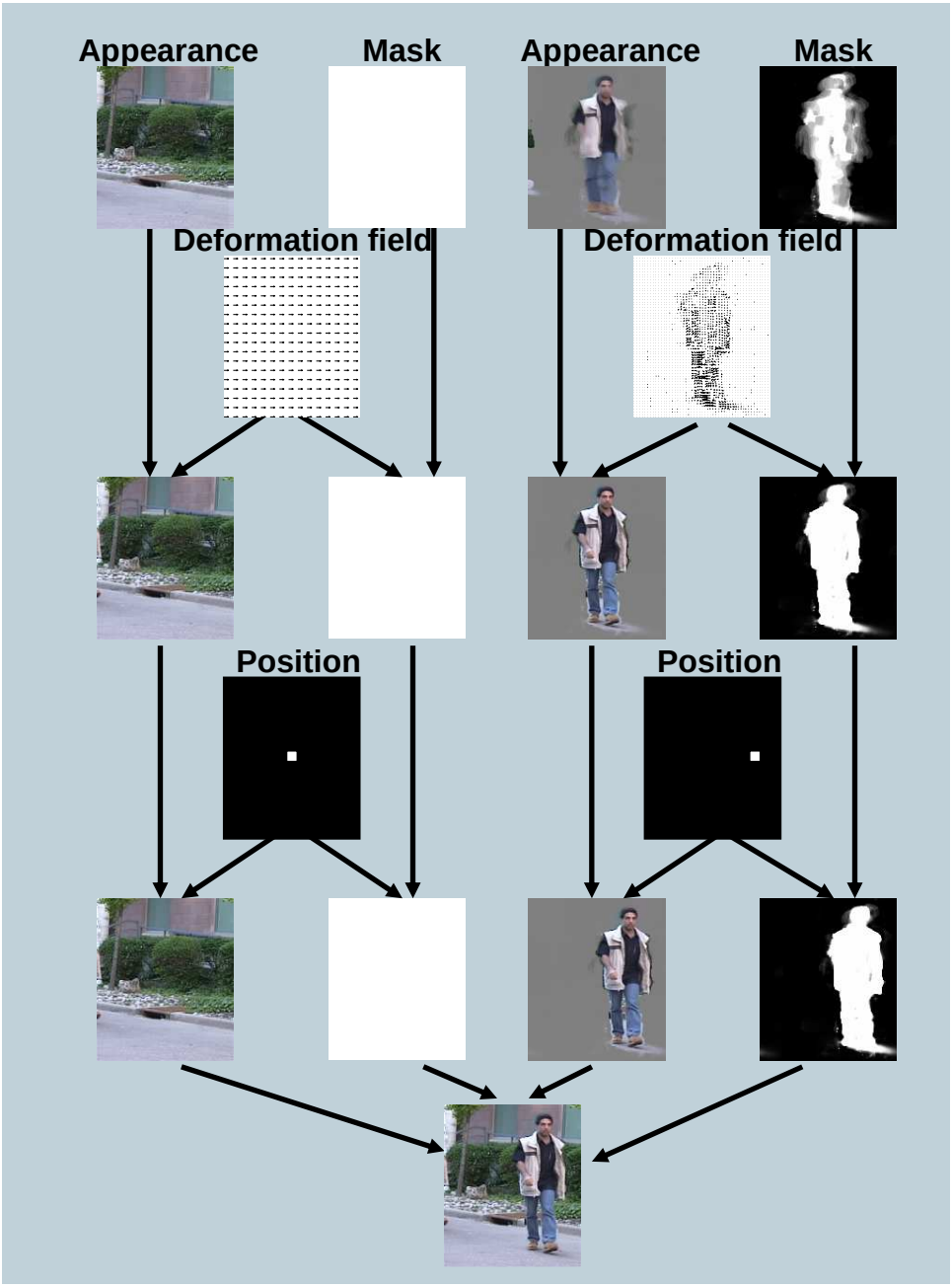

Figure 1: Illustration of the generative model using two layers. The images used in this figure are obtained by learning the model using a video sequence of a person walking towards the camera diagonally. The deformation model is nonlinear and is fully expressive; the additional level of global transformation provides regularization and offers significant computational advantage as discussed in sec. 3, but could be instead absorbed into the deformation field.

Every $M \times N$ observed frame is decomposed into a $(M-K+1)\times(N-K+1)$ grid of $K \times K$ *overlapping* patches in the spirit similar to our epitomic representations (Jojic, Frey and Kannan 2003). We let $\mathcal{P}(\mathbf{z})$ denote the set of coordinates that are in the patch centered at $\mathbf{z}$ so that $\mathcal{P}(\mathbf{z}) = \{\mathbf{w} : |\mathbf{w}-\mathbf{z}| \leq K\}$ and the corresponding pixel intensities to be $I(\mathcal{P}(\mathbf{z}))$

For pixel $I^t(\mathbf{z})$ in the observed image at time t, the foreground motion vector is represented by the random variable $\mathbf{U}^t(\mathbf{z})$, and the background motion vector by the random variable $\mathbf{V}^t(\mathbf{z})$.

To generate a pixel $I^t(\mathbf{z})$ in the observed image at time t, a foreground motion vector , $\mathbf{U}^t(\mathbf{z}) = \mathbf{u}$ from $P(\mathbf{U})$ and a background motion vector $\mathbf{V}^t(\mathbf{z}) = \mathbf{v}$ from $P(\mathbf{V})$ are drawn. The intensity of the pixel in the patch $\mathcal{P}(\mathbf{z})$ at time t is generated using :

$$I^t(\mathbf{w} \in \mathcal{P}(\mathbf{z})) = \mathbf{f}(\mathbf{w}+\mathbf{u})^{\mathbf{m}(\mathbf{w}+\mathbf{u})} * \mathbf{b}(\mathbf{w}+\mathbf{v})^{(1-\mathbf{m}(\mathbf{w}+\mathbf{u}))} + noise$$

Thus, when $\mathbf{m}(\mathbf{w}+\mathbf{u}) = 1$, foreground pixel intensity is observed, and when $\mathbf{m}(\mathbf{w}+\mathbf{u}) = 0$, background pixel intensity is observed at pixel location $\mathbf{w} \in \mathcal{P}(\mathbf{z})$. We assume that the (sensor) noise is Gaussian with variance σ^2 so that the observation likelihood of the patch is a Gaussian given by,

$$P(I^t(\mathcal{P}(\mathbf{z}))|\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) \propto exp\Big[- \sum_{\mathbf{w}\in\mathcal{P}(\mathbf{z})} \frac{(\mathbf{f}(\mathbf{w}+\mathbf{u})^{\mathbf{m}(\mathbf{w}+\mathbf{u})}\mathbf{b}(\mathbf{w}+\mathbf{v})^{1-\mathbf{m}(\mathbf{w}+\mathbf{u})} - I^t(\mathbf{w}))^2}{2\sigma^2}\Big]$$

As in the epitome representation, we assume that the patch appearances are independent.

Let the motion fields in all nearby frames be $\mathcal{U}$ and $\mathcal{V}$ and the observed patches in all nearby frames be $\mathcal{I}$ so that joint distribution is proportional to

$$P(\mathcal{U}, \mathcal{V}, \mathcal{I}, \mathbf{m}) \propto \prod_t \prod_{\mathbf{z}} P(\mathbf{m}) P(I^t(\mathcal{P}(\mathbf{z}))|\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}, \mathbf{m}) \quad (1)$$

5 Inference & Learning

For learning, the natural choice is the Expectation Maximization algorithm (Dempster, Laird and Rubin 1977) that maximizes the likelihood of observation. However, as exact inference is intractable, we resort to variational approximation (Jordan et al. 1999) for the posterior and use generalized EM (Neal and Hinton 1998) for learning.

For each observed image, we approximate posterior as:

$$P(\mathbf{U}, \mathbf{V}, \mathbf{m}|I^t) = \prod_{\mathbf{z}} q^t(\mathbf{U}(\mathbf{z}), \mathbf{V}(\mathbf{z})) q^t(\mathbf{m})$$

Letting β_i^t be the probability that $\mathbf{m}_i$=1 given the i^{th} pixel in the t^{th} frame, and $\overline{\beta_i^t}$ = 1-β_i^t,

$$q^t(\mathbf{m}) = \prod_i (\beta_i^t)^{\mathbf{m}_i} \overline{\beta_i^t}^{(1-\mathbf{m}_i)}$$

Generalized EM maximizes the bound on the (log) probability of the data:

$$logP(\mathcal{I}) \geq \sum_t \sum_{\mathbf{z}} \sum_m \sum_{u,v} q^t(\mathbf{U}(\mathbf{z}), \mathbf{V}(\mathbf{z})) q^t(\mathbf{m}) log \frac{P(U(\mathbf{z}), V(\mathbf{z}), I^t(\mathcal{P}(\mathbf{z})), \mathbf{m})}{q^t(\mathbf{U}(\mathbf{z}), \mathbf{V}(\mathbf{z})) q^t(\mathbf{m})}$$

Before, we derive the update equations, we define two quantities: We allow every patch to shift by at most D pixels. This reduces the search space and therefore the computational cost. When there is large motion, we can further reduce this search space D by incorporating global transformation, as described in sec. 3.

The set of all coordinates in the observed image whose $K \times K$ patches can "reach" coordinate $\mathbf{x}$ in $\mathbf{f}$ or $\mathbf{b}$ when moved by at most D is $\mathcal{R}(\mathbf{x}) = \{\mathbf{z} : |\mathbf{x}-\mathbf{z}| \leq (K-1)/2 + D\}$ The set of all motion vectors for the patch at $\mathbf{z}$ in observed image that cause a pixel in the patch to be mapped to $\mathbf{x}$ in $\mathbf{f}$ or $\mathbf{b}$ is

$$\mathcal{M}(\mathbf{x}, \mathbf{z}) = \{\mathbf{u} : |(\mathbf{x}-\mathbf{z})-\mathbf{u}| \leq (K-1)/2; |\mathbf{u}| \leq D\}.$$

The posterior distribution over the motion vectors is

$$q^t(\mathbf{U}(\mathbf{z})=\mathbf{u}, \mathbf{V}(\mathbf{z})=\mathbf{v}) = \rho \exp\Big[-\frac{1}{2\sigma^2} \sum_{\mathbf{w}\in\mathcal{P}(\mathbf{z})} \Big\{\beta_{\mathbf{w}}^t (I^t(\mathbf{w}) - \mathbf{f}(\mathbf{w}+\mathbf{u}))^2 + \overline{\beta_{\mathbf{w}}^t} (I^t(\mathbf{w}) - \mathbf{b}(\mathbf{w}+\mathbf{v}))^2\Big\}\Big] \quad (2)$$

where ρ ensures that $\sum_{\mathbf{u}} \sum_{\mathbf{v}} P(\mathbf{U}^t(\mathbf{z}) = \mathbf{u}, \mathbf{V}^t(\mathbf{z}) = \mathbf{v}|I^t) = 1$. Due to the use of binary mask, the computation inside the exponential splits into sum of two distance measures. When the posterior over the mask is peaked, pixels in the observed patch that are attributed to foreground are compared with patch from the shifted foreground, and observed pixels that belong to background are matched with the corresponding shifted patch in the background. This distance computation need not be done on a patch by patch basis, but instead by observing that each pixel participates in a large number of patches, we can employ simple trick using cumulative sums and calculate the distances for all patches in tandem.

The posterior distribution over the mask is

$$\beta_{\mathbf{w}}^t = 1/\Big[1 + exp\Big(\sum_{\mathbf{u},\mathbf{v}} q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) ((I^t(\mathbf{w}) - \mathbf{b}(\mathbf{w}+\mathbf{v}))^2 - (I^t(\mathbf{w}) - \mathbf{f}(\mathbf{w}+\mathbf{u}))^2)\Big)\Big] \quad (3)$$

Figure 2: Top row: entire sequence of 6 frames used to train the model. Notice that one person is moving towards the camera, inducing a zooming in effect, and another person moving in the background. Bottom row: interpolated frame between adjacent frames in the top row. Interpolation is performed using the learned parameters and the inferred deformation fields.

We use $< . >= \sum_t \sum_{\mathbf{z}\in\mathcal{R}(\mathbf{y})} \sum_{\mathbf{u}\in\mathcal{M}(\mathbf{y},\mathbf{z})} \sum_{\mathbf{v}\in\mathcal{M}(\mathbf{y},\mathbf{z})}$ to represent the sufficient statistic collected from all pixels, $\mathbf{z}$ in all frames, t, that map pixel $\mathbf{y}$ and the motion vectors for $\mathbf{z}$ that cause the pixel to be mapped.

The update for background appearance is:

$$\mathbf{b}(\mathbf{y}) \leftarrow \frac{\left\langle \overline{\beta^t_{\mathbf{y}-\mathbf{v}}} q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) I^t(\mathbf{y}-\mathbf{v}) \right\rangle}{\left\langle \overline{\beta^t_{\mathbf{y}-\mathbf{v}}} q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) \right\rangle} \quad (4)$$

The above update involves aligning the observed pixel with respect to the background using the posterior distribution over the motion vectors and then multiplying this with the posterior probability of the pixel belonging to the background. Since multiple patches from all the frames contribute to updating the same pixel in the background, the denominator normalizes for multiple counts.

The foreground appearance is updated similarly:

$$\mathbf{f}(\mathbf{y}) \leftarrow \frac{\left\langle \beta^t_{\mathbf{y}-\mathbf{u}} q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) I^t(\mathbf{y}-\mathbf{u}) \right\rangle}{\left\langle \beta^t_{\mathbf{y}-\mathbf{u}} q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) \right\rangle} \quad (5)$$

The prior probability of a pixel to be from the foreground is given by:

$$\alpha_{\mathbf{y}} \leftarrow \frac{\left\langle \beta^t_{\mathbf{y}-\mathbf{u}} q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) \right\rangle}{\left\langle q(\mathbf{U}^t(\mathbf{z})=\mathbf{u}, \mathbf{V}^t(\mathbf{z})=\mathbf{v}) \right\rangle} \quad (6)$$

We initialize the appearance variables to a reference frame (usually the middle frame in the sequence) and let the prior distribution over the mask to be uniform. We iterate between finding the posterior over motion vectors and the posterior over the mask in the Estep, and updating the appearances and the prior distribution for the mask in the Mstep. This procedure enables inferring the layered optical flow in images. However, in the full flexible sprites model, the chosen variational factorization, when combined with the variational factorization of the shifts in the original flexible sprites paper leads to efficient inference and learning whose results are shown in Fig. 1. We omit the mathematical details for brevity.

6 Experimental results

6.1 Modelling complex deformation and appearances in two layers

For this experiment, we used 6 frames of 88×133 RGB sequence shown in fig. 2. The sequence has a person moving towards the camera in front of a moving background inducing a complex deformation field. There is also another person walking behind in the opposite direction. The translational motion of the background is due to camera shake.

We trained our foreground-background model on this sequence using 5×5 overlapping patches. For computational reasons, we restricted the search space for the foreground motion to be 4 pixels in both directions (81 possible directions). The background motion was restricted to 2 pixels in the horizontal direction. The state space of the posterior distribution over the motion field has a cardinality of 405 ($9 \times 9 \times 5 \times 1$). Upon investigation, we found that the posterior distribution is peaked at a few values. This fact can be used to address the storage issue during inference. In fact, in our experiments we store the distribution of only the top 20 motion directions.

In Fig. 3b, we show the learned appearances of the foreground and background, and the probability distribution of the binary mask. It is interesting to notice that the learned mask distribution has captured the person walking behind as part of the foreground. Also, some of the occluded background pixels are filled in. In fig. 3a, results of learning flexible sprites model without deformation is shown. Since, the flexible sprites

Figure 3: Top row: *Flexible sprites model* Background, foreground (masked) and transparency mask learned using the two layer sprites model on the data in Fig. 2. The complex deformation of the foreground object can't be modelled using this model.
Bottom row:*Proposed model* Learned background, foreground and probability values of binary mask learned. The foreground appearance is invariant to deformation, and the background has lesser number of occluded pixels, and the mask captures the person moving away from the camera as part of the foreground.

model assumes that the pixels belonging to a layer move with the same velocity it can not handle non-uniform motion. In fact, some of the foreground pixels are misclassified to be background pixels as the corresponding background pixels are always occluded.

Inference in this model also gives us the distribution over the motion vectors for each pixel and for each layer. Using this we can compute the expected motion for each pixel by averaging the foreground and background motion weighted by their posterior probabilities. The middle frame is considered as the reference frame for which the motion is set to be 0. Fig. 4 shows the inferred deformation field for each frame with respect to the reference frame. Note, however, that our motion field is defined with respect to the *derived* foreground and background appearances in $\mathbf{b}$ and $\mathbf{f}$ which have more disoccluded pixels than any frame.

The learned appearances and the inferred flow vectors can be used to perform video interpolation. In Fig. 2 we present 1 frame interpolation between adjacent frames. See the accompanying website for more interpolation results.

6.2 Modelling mixtures of complex deformation and appearances in two layers

Our model can easily be extended to incorporate multiple layers of moving objects with appearance of each layer modelled as a mixture model.

In this experiment, we present results for learning a two layer model where the foreground appearance is modelled using a Gaussian mixture with 2 classes. We also allow the latent variables (appearances and probability masks) to be bigger than the observed image so as to learn a panoramic background.

We learn the model using 10 RGB frames ($138 \times 148 \times 3$) sampled from a longer video sequence (Fig. 6a). Each frame consists of one of the two persons (modelled using different classes) moving towards the camera in front of a non-stationary background. Notice that the images include scale changes in appearance due to zoom, complex motion of hands and legs, wrinkles in the clothing, and large shifts in the position.

We used larger appearances and masks (138×178) than the observed frames. Referring to our model in fig. 1, we first train the model without incorporation of the deformation to obtain the global position variables in each layer for each frame. Once the global shifts are inferred, we fix them to learn the deformation field and the parameters of the model in tandem as outlined in Sec. 5

In fig. 5, the parameters of the learned model are shown. Frames corresponding to the first appearance class have pixels belonging to the background that are always occluded. However, these pixels are visible in some frames where the other appearance class is present. By jointly modelling all the frames, we are able to fill in for almost all the occluded pixels belonging to the background for any given frame. This is further shown in fig. 5a. If we had chosen to learn two separate models for the two classes, the background will not have all its corresponding pixels observed.

For the pixels belonging to the texture less pathway, the prior probability distribution over the mask is close to uniform (fig. 5c & e). This suggests that for those regions that do not have enough textural variations to group them as belonging to one of the two layers, it is at best to assign equal probability for either layer to explain them.

In fig. 6, we present inference results for some representative frames, shown in fig. 6a. Fig. 6b is the corresponding inferred deformation field shifted according to inferred global transformation. The flow vectors

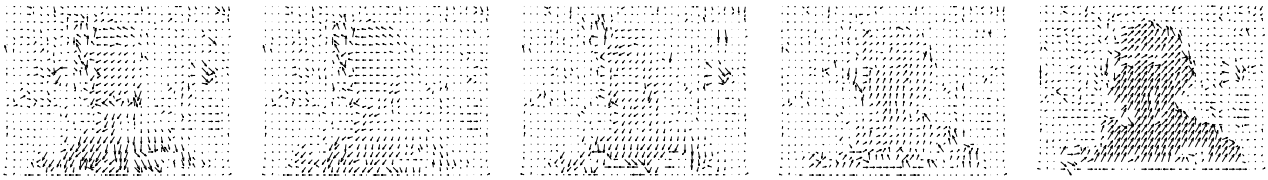

Figure 4: Inferred deformation field corresponding to the image sequence shown in Fig. 2(the flow field is drawn with reference to the fourth frame which is not shown here)

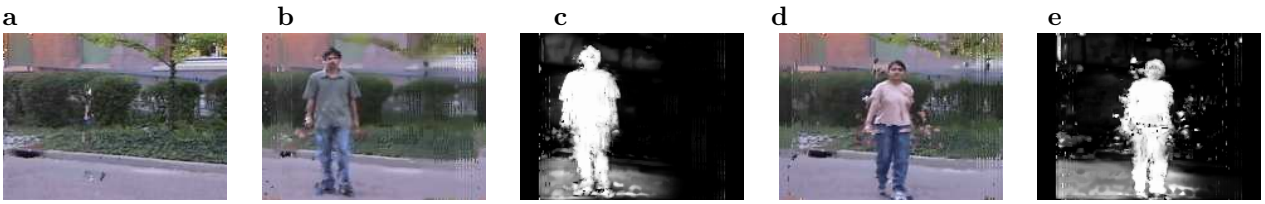

Figure 5: Parameters of two layered two class model learned using 10 frames from a video sequence (representative frames in fig. 6 a)Learned background is larger than the size of input image b) Appearance of foreground object of class 1 and c) the corresponding probability distribution of the binary mask (with white referring to probability of 1 for the pixel belonging to foreground) d) & e) appearance and probability mask of the second class of foreground layer.

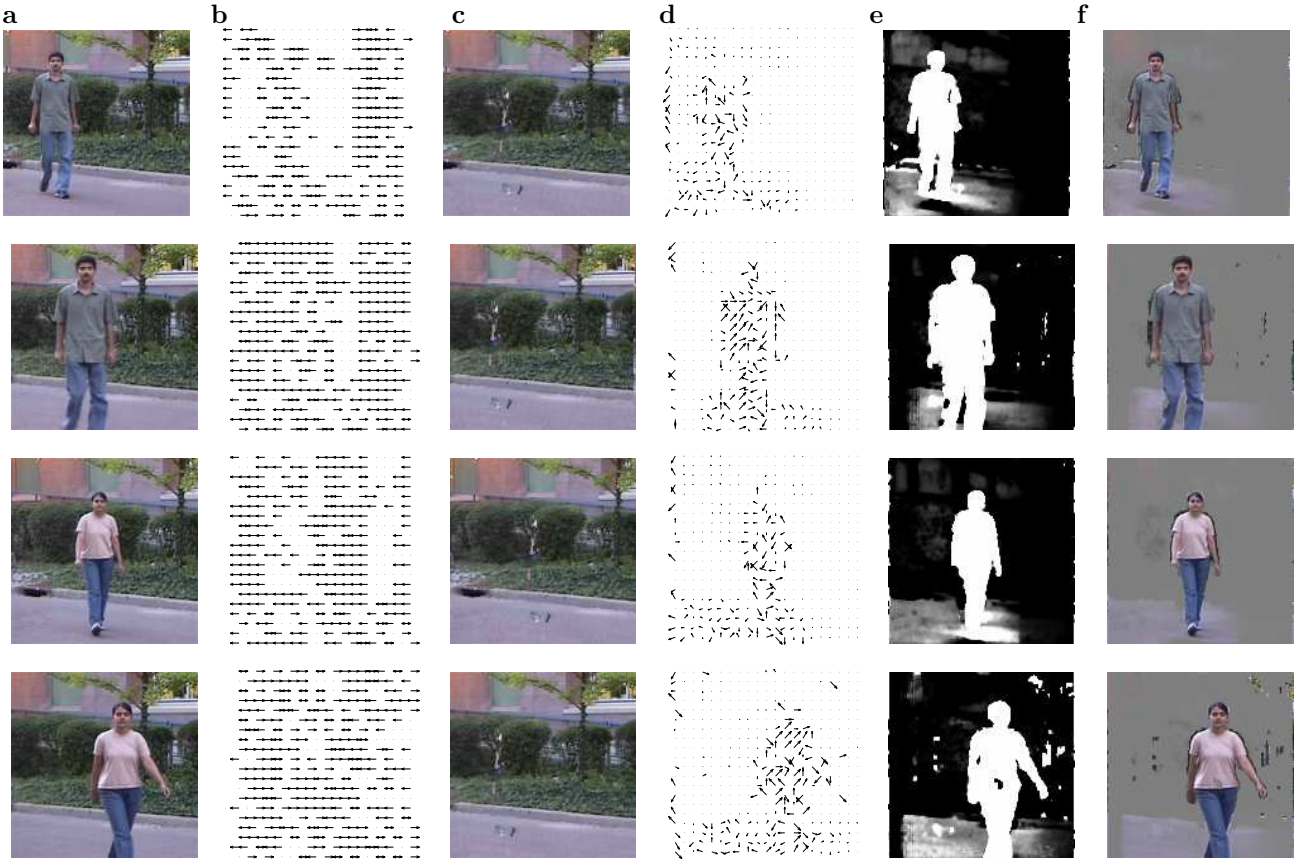

Figure 6: Illustration of inference for some frames of the sequence explained in sec. 6.2 a) frames from a sequence. b) Deformation field for the background. c) Deformed, globally transformed background. d) Deformation field for the foreground (masked). e) Distribution for the mask after global transformation is applied. f) Mask applied on the frame in a)

are drawn relative to the learned parameters. Each vector in this field represents the most probable (that vector that has the largest posterior probability) deformation vector for that pixel. The inferred deformation field for the foreground appearance is in fig. 6c. It is interesting to note that the deformation vectors for appearance are smoother and more consistent along the boundaries than on the central regions of the foreground object. As our approach learns a good appearance model, the boundaries of the objects are well defined but the motion within the object is not very coherent due to lack of enough texture variation between adjacent patches to reliably favor a particular motion direction. We contrast this with inferred flow vectors in the previous experiment (fig. 4) where we had enough textural variation in the central region of object of interest that we learned a much smoother flow field.

7 Conclusions

We have enriched the flexible sprites model with the deformable motion variables defined on overlapping patches. We assume that in each patch there exist two motion vectors and that some pixels are following one and others the other motion. The selection is defined by a patch of binary variables. These patches are also overlapping in the model of the mask, aligned with one of the layers. We were able to use this model of motion within the flexible sprites model and obtain better appearance, mask and motion estimates. See the web page http://www.psi.utoronto.ca/∼anitha/flex.html for additional results and videos.

Acknowledgments

The authors thank P.Anandan for discussion on the importance of combining top-down object appearance models with low-level visual cues, in particular, motion. We also thank Allan Jepson for his comments on an earlier version of the work.

References

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

Black, M. J. and Fleet, D. J. 2000. Probabilistic detection and tracking of motion discontinuities. *International Journal on Computer Vision*.

Black, M. J. and Jepson, A. 1996. Estimating optical flow in segmented images using variable-order parametric models with local deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Proceedings of the Royal Statistical Society*, B-39:1–38.

Frey, B. and Jojic, N. 2004. Advances in algorithms for inference and learning in complex probability models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (*to appear*)*.

Frey, B. J. and Jojic, N. 1999. Estimating mixture models of images and inferring spatial transformations using the em algorithm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE.

Frey, B. J., Jojic, N., and Kannan, A. 2003. Learning appearance and transparency manifolds of occluded objects in layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jepson, A. and Black, M. J. 1993. Mixture models for optical flow computation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 760–761.

Jojic, N., Frey, B., and Kannan, A. 2003. Epitomic analysis of appearance and shape. In *Proceedings of International Conference in Computer Vision*. IEEE.

Jojic, N. and Frey, B. J. 2001. Learning flexible sprites in video layers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jojic, N., Simard, P., Frey, B. J., and Heckerman, D. 2001. Separating appearance from deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. 1999. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233.

Lucas, B. D. and Kanade, T. 1981. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*.

Neal, R. M. and Hinton, G. E. 1998. A view of the em algorithm that justifies incremental,sparse, and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Publishers, Norwell MA.

Weiss, Y. and Adelson, E. 1996. A unified mixture framework for motion segmentation: Incorporating spatial coherence and estimating the number of models. In *Proceedings of IEEE Computer Vision and Pattern Recognition*.

Williams, C. W. and Titsias, M. K. 2003. Learning about multiple objects in images: Factorial learning without factorial search. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*. MIT Press, Cambridge MA.