Explaining Optical Flow Events with Parameterized Spatio-temporal Models

Michael J. Black *

Xerox Palo Alto Research Center
3333 Coyote Hill Road, Palo Alto, CA 94304
black@parc.xerox.com, http://www.parc.xerox.com/black/

Abstract

A spatio-temporal representation for complex optical flow events is developed that generalizes traditional parameterized motion models (e.g. affine). These generative spatio-temporal models may be non-linear or stochastic and are event-specific in that they characterize a particular type of object motion (e.g. sitting or walking). Within a Bayesian framework we seek the appropriate model, phase, rate, spatial position, and scale to account for the image variation. The posterior distribution over this parameter space conditioned on image measurements is typically non-Gaussian. The distribution is represented using factored sampling and is predicted and updated over time using the Condensation algorithm. The resulting framework automatically detects, localizes, and recognizes motion events.

1 Introduction

While the last decade has seen significant improvements in the robustness and accuracy of techniques for estimating optical flow, a number of issues remain. Consider the optical flow field computed for the pair of images in Figure 1 [13]. The recovered flow serves as a poor characterization of the image motion but illustrates two open problems. First, motion remains difficult to estimate in complex image sequences containing non-rigid deformations, articulated motions, occlusion, illumination changes, self shadowing, etc. Second, even if we could estimate the motion accurately, how can the optical flow be used to recognize the activity? In contrast to traditional optical flow methods this paper presents a framework that shifts the focus of the problem from *estimation* of accurate pixel motion in generic scenes to *explanation* of image changes in terms of explicit spatio-temporal models of motion events.

These spatio-temporal models are illustrated in Figure 2. The spatial component consists of a basis set of flow fields, $\vec{\mathbf{b}}_j$. The temporal component, $\vec{\tau}_k$, contains trajectories of coefficients. A particular position, or phase ϕ, within the temporal model determines a vector of linear coefficients, $\vec{\mathbf{a}}_{k,\phi} = [a_1, \ldots, a_n]$ that, together with the spatial model determines a flow field. Particular motion events are described by a spatio-temporal model $\mu_i = \{\vec{\mathbf{b}}_j, \vec{\tau}_k\}$.

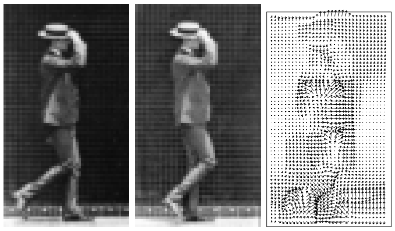

Figure 1: Challenges for motion *estimation*/*explanation*.

Recognizing a motion event requires choosing the most likely spatio-temporal model, μ, the correct phase, the position, $\vec{\mathbf{p}}$, of the model within the image, the spatial scale, ζ, and an amplitude scaling, α. An additional rate parameter ρ will described later. With multiple, non-linear, models, the probability distribution over these parameters is non-Gaussian and we represent it explicitly using a discrete set of random samples.

This distribution can be predicted forward in time and updated with new information using the Condensation algorithm [8]. With a high dimensional parameter space, many samples may be needed to characterize the distribution and each sample requires computing the likelihood of the image measurements given a particular set of parameters. This can be done efficiently by sampling from spatial and temporal image derivatives in a multi-scale representation. Note the likelihood function is computed directly from image derivatives and not from a dense optical flow field.

The resulting model solves a number of problems simultaneously. Motion events are detected, localized, and recognized automatically using only motion information. The approach extends the application of parameterized spatial models to domains that require non-linear, or stochastic, spatio-temporal models. The models provide strong constraints on the interpretation of motion which are exploited to estimate motion in challenging sequences such as that in Figure 1. More importantly however, the approach shifts the focus from accurate estimation of general flow fields to

* Portions of this work were performed while the author was on sabbatical in the Computational Vision and Active Perception Laboratory (CVAP), Royal Instititue of Technology (KTH), Stockholm, Sweden.

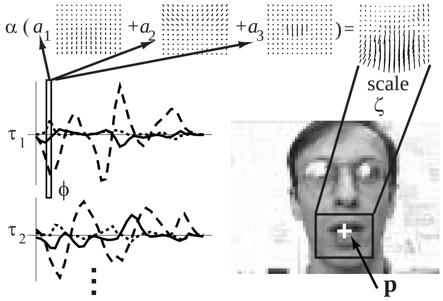

Figure 2: Overview of the generative model.

model-based recognition of motion events.

2 Related Work

Much of the work on detecting and recognizing human motion relies on image brightness [12, 14] or image differences [7] rather than explicitly on optical flow. Unlike motion-based approaches, these methods can be sensitive to the color of the background and the clothing of the subject. They may be most useful in conjunction with a motion-based method as a source of additional information.

Work that uses image motion for recognition typically exploits statistical properties of the motion field [11, 15]. These models allow discrimination between motions with distinct first or second order statistics but cannot model the precise spatio-temporal variation needed to distinguish the mouth motions for two words such as "print" and "track."

Learned models of image motion provide a more precise characterization of the spatial variation of complex objects such as mouths [5]. These spatial models have typically not included models of the temporal properties of the moving objects. Combined spatial and temporal models have typically exploited assumptions of constant acceleration resulting in a linear formulation [18]. Linear temporal models are not sufficient for the recognition of complex motion events.

More complex spatial models of human motion treat the body as a set of connected parts [4, 6, 10]. There has also been work on recognition of activities using the temporal evolution of the parameters of models such as these [3, 16]. Yacoob and Davis [17] use learned spatial models of walking motions to constrain the tracking and estimation of a part-based motion model. All these tracking methods require manual initialization of the spatial model.

To cope with more complex noise models and the problem of initialization, Isard and Blake proposed the Condensation algorithm [8]. By representing a discretely sampled distribution over model parameters they can track objects when multiple matches and occlusion make the distribution non-Gaussian. The approach as been extended to multiple models, multiple sources of information [9], and non-linear temporal models [3]. This paper extends these methods to exploit optical flow information.

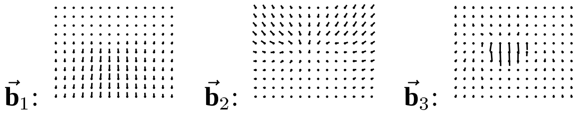

Figure 3: First 3 basis flow fields accounting for 85% of the variance in the mouth training motions.

3 Spatio-temporal Flow Models

This section briefly reviews object specific spatial and temporal models of motion.

Spatial Models of Motion. We formulate spatial models of image motion using a *basis set* of orthogonal flow fields, $\vec{\mathbf{b}}_j$. Linear combinations of these basis flow fields are used to describe image motion. Basis flow fields my be constructed from examples using principal component analysis [5] and may be used to approximate the motions of fairly complex non-rigid objects such as human mouths. While we restrict our attention to linear spatial models here the framework is more general and can be used with non-linear or stochastic models of motion.

As an example, we construct a spatial model of mouth motion for a single speaker. A training set of 3000 images is gathered of a person saying several words and changing facial expressions throughout several seconds of video. The face region is stabilized using an affine motion model [4] and the motion of the mouth region is estimated relative to the stabilized sequence using a dense flow method [2]. Singular value decomposition is used to compute a set of basis flow fields from the training flow fields [5] (Figure 3). A small number of basis flow fields will be sufficient to discriminate between different optical flow events.

Temporal Models of Motion. While spatial models constrain instantaneous motion, the temporal properties of many motions may be modeled to further constrain the interpretation of image brightness changes. Combined spatial and temporal models can be constructed by performing principal component analysis on a space-time block of training flow fields.

For objects such as human mouths, however, we can separate the spatial and temporal models. The spatial variation of the mouth is modeled with the spatial basis flow fields above. Different words or expressions will result in patterns of image motion that are modeled by different discrete trajectories of the spatial coefficients over time. Temporal model k is $\vec{\tau}_k = [\vec{\mathbf{a}}_{k,1}, \ldots, \vec{\mathbf{a}}_{k,\phi_{k,\max}}]$, where $\phi_{k,\max}$ is the length of the model and $\vec{\mathbf{a}}_{k,\phi}$ is a vector of linear coefficients at phase ϕ. A spatio-temporal model $\mu_i = \{\vec{\mathbf{b}}_j, \vec{\tau}_k\}$ combines a spatial basis set $\vec{\mathbf{b}}_j$ and a temporal model $\vec{\tau}_k$.

For example, the mouth training set above contains multiple utterances of the words "Center," "Print," "Track," and "Release;" these words were chosen for a user interface ap-

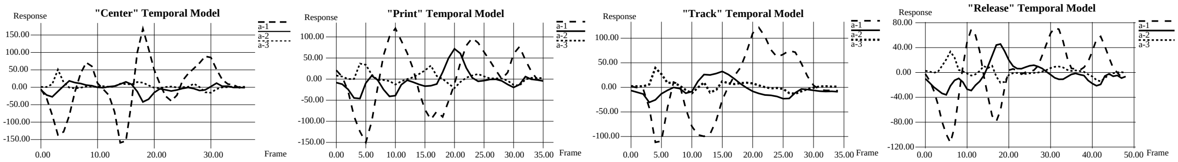

Figure 4: Temporal models for "Center," "Print," "Track," and "Release" utterances; dash = a_1, solid = a_2, dot = a_3.

plication. Given the spatial basis set we estimate the coefficients describing the image motion using the method in [5]. For each utterance we take the trajectories of the motion coefficients and manually align them [3]. The mean trajectories shown in Figure 4 form the temporal models.

4 Model-based Motion Representation

Given an image sequence and a set of spatio-temporal models of motion events, explaining the image in terms of the models requires that we choose a model μ, at spatial location $\vec{\mathbf{p}} = (x, y)$ and scale ζ, with a rate, ρ, an amplitude, α, and a temporal position, or phase, ϕ. This defines a *generative model* of the image motion in a particular region. Note that α and ζ are assumed independent.

Let $\mathbf{s}_t = (\mu, \vec{\mathbf{p}}, \zeta, \rho, \alpha, \phi)$ be a *state* at time t. Let $z_t = [I_x(t), I_y(t), I_t(t)]$ represent the spatial and temporal derivatives of the image sequence at time t and let $\vec{\mathbf{z}}_t$ be a sequence of such derivatives from time 0 to time t. Then we seek the probability of a state, $\mathbf{s}_t$, given the data, $\vec{\mathbf{z}}_t$. This distribution is not directly observable, but from Bayes' rule

$$p(\mathbf{s}_t | \vec{\mathbf{z}}_t) = k\, p(z_t | \mathbf{s}_t)\, p(\mathbf{s}_t | \vec{\mathbf{z}}_{t-1}), \tag{1}$$

where the measurement density $p(z_t | \mathbf{s}_t)$ can be evaluated for a state, $p(\mathbf{s}_t | \vec{\mathbf{z}}_{t-1})$ is the prior probability of a particular state, and k is a normalizing constant independent of $\mathbf{s}_t$.

The non-linear nature of the motion models means that $p(\mathbf{s}_t | \vec{\mathbf{z}}_t)$ will not be Gaussian and we represent this distribution using a discrete set of random samples [8]. A similar representation is used for dense flow estimation in [19].

Measurement Density. We need to compute the likelihood, $p(z_t | \mathbf{s}_t)$ of observing the image measurements given a state. Recall that a state determines a vector of coefficients and hence a flow field. Using this, we define the likelihood in terms of the brightness constancy assumption

$$I(\vec{\mathbf{x}}(\mathbf{s}_t), t) = I(\vec{\mathbf{x}}(\mathbf{s}_t) + \vec{\mathbf{u}}(\vec{\mathbf{x}}; \mathbf{s}_t), t-1)$$

which states that the image, I, at time t is a warped version of the image at time $t-1$. The spatial position $\vec{\mathbf{x}}(\mathbf{s}_t)$ takes into account the spatial scale ζ and transforms the image to the scale of the spatial basis flow fields

$$\vec{\mathbf{x}}(\mathbf{s}_t) = \lfloor \zeta(\vec{\mathbf{x}} - \vec{\mathbf{p}}) + \vec{\mathbf{p}} \rfloor,$$

where $\vec{\mathbf{x}}$ is a position in image coordinates and $\vec{\mathbf{p}}$ is the location of $\mathbf{s}_t$ (with respect to the center of the spatial basis) in image coordinates.

The optical flow at a pixel $\vec{\mathbf{x}}$ is

$$\vec{\mathbf{u}}(\vec{\mathbf{x}}; \mathbf{s}_t, \hat{t}\,) = \sum_{i=1}^{n} \hat{a}_i(\phi(\hat{t}; \mathbf{s}_t)) \vec{\mathbf{b}}_{\mu,i}(\vec{\mathbf{x}}(\mathbf{s}_t) - \vec{\mathbf{p}}),$$

where $\vec{\mathbf{b}}_{\mu,i}(\vec{\mathbf{x}}(\mathbf{s}_t) - \vec{\mathbf{p}})$ is the basis flow field i for the event model μ. Typically, $\hat{t} = t$ but we can evaluate the phase at some time $\hat{t} < t$ in which case the rate parameter ρ is used to compute the phase: $\phi(\hat{t}; \mathbf{s}_t) = \phi - \rho(t - \hat{t}\,)$.

Temporal trajectories are represented by coefficient values at discrete time instants. To compute the coefficient $\hat{a}_i(\phi)$ for some real ϕ, $0 \leq \phi < \phi_{\max}$, we linearly interpolate the coefficients and scale the result by α to allow variations in amplitude

$$\hat{a}_i(\phi) = \alpha(a_i(\lfloor\phi\rfloor)(1 - (\phi - \lfloor\phi\rfloor)) + a_i(\lfloor\phi\rfloor + 1)(\phi - \lfloor\phi\rfloor)).$$

For computational efficiency we linearize the brightness constancy assumption about $\vec{\mathbf{u}}(\vec{\mathbf{x}}; \mathbf{s}_t)$ to derive the optical flow constraint equation at a pixel. We use a Gaussian pyramid image representation and compute image derivatives at all scales. The constraint equation at a particular pyramid level g is

$$\nabla I(\vec{\mathbf{x}}(\mathbf{s}_t), g, \hat{t}\,)\vec{\mathbf{u}}(\vec{\mathbf{x}}; g, \mathbf{s}_t, \hat{t}\,) + I_t(\vec{\mathbf{x}}(\mathbf{s}_t), g, \hat{t}\,) = E(\vec{\mathbf{x}}, \mathbf{s}_t, g, \hat{t}\,)$$

where ∇I represents a vector of spatial derivatives $[I_x, I_y]$ at time $\hat{t}$. Note that we also have a pyramid of basis flow fields corresponding to the spatial scales in the image pyramid.

To evaluate the likelihood $p(z_t | \mathbf{s}_t)$ we take a random sample, $\mathcal{R}_g$, of image locations at each level g (typically 1% of the pixels in the region) (cf [1]) and use the error above to define $p(z_t | \mathbf{s}_t) =$

$$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2\sigma^2 \sum_{\hat{t},g} |\mathcal{R}_g|} \sum_{\hat{t}=t}^{t-w} \sum_{g} \sum_{\vec{\mathbf{x}} \in \mathcal{R}_g} \tilde{E}(\vec{\mathbf{x}}, \mathbf{s}_t, g, \hat{t}\,)^2\right)$$

where $\tilde{E}(\vec{\mathbf{x}}, \mathbf{s}_t, g, \hat{t}) = \min(|E(\vec{\mathbf{x}}, \mathbf{s}_t, g, \hat{t})|, \sigma_{\text{Out}})$, and σ_{Out} is the residual of an outlier (empirically chosen to be 15.0). Note that the likelihood is computed over a temporal window, w, (taken to be 5 frames). This violates the assumption in (1) that observations are mutually conditionally independent but allows time warping in the computation of the likelihood and reduces the number of discrete samples needed to represent the posterior.

The value of $\hat{a}(\phi)$ when $\phi < 0$ depends on the model μ. In the case of repetitive actions such as walking the phase is computed modulo the $\phi_{\max}$. In the case of mouth motions we assume the $\hat{a}_i(\phi) = 0$ when $\phi < 0$; that is the mouth is static before an utterance. Position, $\vec{\mathbf{p}}$, can also depend on $\hat{t}$ but here we assume it is constant over w.

The distribution $p(\mathbf{s}_t|\vec{\mathbf{z}}_t)$ will be represented below with discrete samples. Each sample requires evaluating the likelihood $p(z_t|\mathbf{s}_t)$ and hence this computation must be efficient. For this reason, we place a number of restrictions on the samples $\mathcal{R}_g$. First, we support non-rectangular basis flow models by restricting the samples $\mathcal{R}$ to be drawn from a binary mask distribution that indicates where the basis flow model is defined. Second, we restrict the samples $\vec{\mathbf{x}} \in \mathcal{R}_g$ to be those for which the generated flow $\vec{\mathbf{u}}(\vec{\mathbf{x}}; g, \mathbf{s}_t, \hat{t}\,)$ magnitude in the horizontal and vertical direction is less than 1.5 pixels; this restricts large motions to being evaluated only at the appropriate scales, g. Third, so that all constraints are as informative as possible, we restrict the samples to locations where the spatial brightness variation (∇I) is greater than a threshold; this reduces the required number of samples. Given these constraints on $\mathcal{R}$, if sufficient samples are not available to evaluate the likelihood, the sample is assigned the probability of an outlier.

Temporal Prior. The prior $p(\mathbf{s}_t|\vec{\mathbf{z}}_{t-1})$ embodies the temporal dynamics of the system

$$p(\mathbf{s}_t|\vec{\mathbf{z}}_{t-1}) = \sum_{i=1}^{S} p(\mathbf{s}_t|\mathbf{s}_{t-1}^{(i)})\, p(\mathbf{s}_{t-1}^{(i)}|\vec{\mathbf{z}}_{t-1})$$

where $p(\mathbf{s}_{t-1}^{(i)}|\vec{\mathbf{z}}_{t-1})$ is the posterior from the previous time step. The term $p(\mathbf{s}_t|\mathbf{s}_{t-1}^{(i)})$ defines how a state evolves over time and will be defined below.

Additional Evidence. The state space we need to represent is large and it is useful to have additional information so that a small number of samples can adequately characterize it. Isard and Blake [9] describe the technique of importance sampling for incorporating additional information when this information is not conditionally independent of the evidence $\vec{\mathbf{z}}_t$.

As in [9], we are concerned with narrowing the search over spatial positions $\vec{\mathbf{p}}$. In some cases we have evidence for a particular location from a higher level model; for example, leg locations given knowledge of the torso position. Alternatively, evidence may come from some image source; for example, mouth locations using color information. In these cases it is reasonable to assume that the evidence is conditionally independent of the image derivatives used to compute the likelihood above.

With additional evidence, m_t, the posterior is

$$p(\mathbf{s}_t|\vec{\mathbf{z}}_t, m_t) = k\, p(z_t|\mathbf{s}_t)\, p(m_t|\mathbf{s}_t)\, p(\mathbf{s}_t|\vec{\mathbf{z}}_{t-1}).$$

Initialization Prior. When initializing a new state with no evidence, let $\mathbf{s}_0$ represent some unknown previous state and $p(\mathbf{s}_t|\mathbf{s}_0)$ be an initialization prior. This distribution is uniform (between minimum and maximum values) over the state parameters μ, α, ρ, ζ. The choice of ϕ depends on the model and may be chosen uniformly or, for mouth motions we chose y uniformly between 0 and 1 and let $\phi = (1 - \sqrt{y})/\sqrt{y}$ which biases new states to have a value of ϕ close to zero. We chose the location $\vec{\mathbf{p}}$ by sampling from $p(m_t|\mathbf{s}_t)$ if the evidence m_t is available; uniformly otherwise.

5 Computational Model

Due to the non-Gaussian nature of $p(z_t|\mathbf{s}_t)\, p(m_t|\mathbf{s}_t)$ we represent this distribution using a finite set of samples, S, [8] and normalize the probabilities of the samples so that they sum to one, producing weights $\pi_t^{(n)}$

$$\pi_t^{(n)} = \frac{p(z_t|\mathbf{s}_t^{(n)})p(m_t|\mathbf{s}_t^{(n)})}{\sum_{i=1}^{S} p(z_t|\mathbf{s}_t^{(i)})p(m_t|\mathbf{s}_t^{(i)})}.$$

The set of S pairs, $(\mathbf{s}_t^{(n)}, \pi_t^{(n)})$, represents the distribution.

Note that given the sampled approximation to the distribution $p(\mathbf{s}_t|\vec{\mathbf{z}}_t)$, we can compute the expected value for some state parameter, $f(\mathbf{s}_t)$, as

$$E[f(\mathbf{s}_t)|\vec{\mathbf{z}}_t, m_t] = \sum_{n=1}^{S} f(\mathbf{s}_t^{(n)})\pi_t^{(n)}.$$

To approximate the prior at time t, we sample from the posterior from time $t-1$ by choosing a state, $\mathbf{s}_{t-1}^{(n)}$ according to the weights $\pi_{t-1}^{(n)}$. To avoid becoming trapped in local maxima we chose some fraction $0 \leq q \leq 1$ of the samples from the initialization prior (q is typically 0.1).

Dynamical Model. Given a sampled state $\mathbf{s}_{t-1}^{(n)}$ we predict the parameters of the new state $\mathbf{s}_t^{(n)}$ at time t by sampling from $p(\mathbf{s}_t|\mathbf{s}_{t-1})$. We assume the model does not change ($\mu_t = \mu_{t-1}$), that ϕ_t is normally distributed about $\phi_{t-1} + \rho_{t-1}$, and that α_t, ρ_t, ζ_t, and $\vec{\mathbf{p}}_t$ are all normally distributed about their values at $t-1$. The normal distribution about $\vec{\mathbf{p}}$ allows the estimated position to drift to follow small changes in location; the temporal dynamics could be extended to allow constant velocity or acceleration.

Given the new states, we evaluate their probability with respect to the evidence at time t and normalize to update the weights. The Condensation algorithm [8] repeats this sampling, prediction, and updating to propagate the posterior over time.

6 Experimental Results

The framework is tested below on walking and speaking events. The number of samples, S, is taken to be 35000.

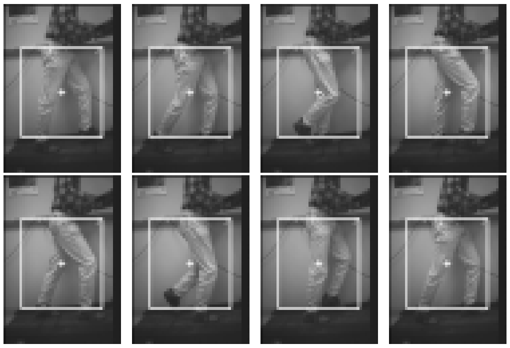

Figure 5: Example training images showing one complete walking cycle. The white box indicates the region of the image used for training.

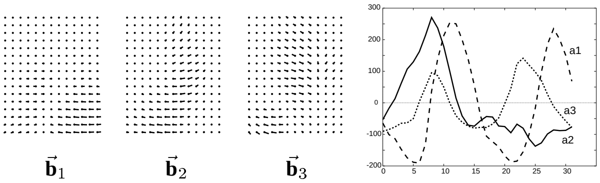

Figure 6: Left: First 3 basis flow fields for walking motion; they account of 80% of the variance in the training motions. Right: temporal model.

6.1 Walking Experiments

We construct spatial and temporal models (Figure 6) from a training sequence of a single subject walking on a treadmill parallel to the image plane of the camera (Figure 5). Note that the temporal model includes the full walking cycle and that each half of the cycle is similar, resulting in phase ambiguities. We assume that the torso location and scale are known and these predict $\vec{\mathbf{p}}$ and ζ and that $p(m_t|\mathbf{s}_t)$ is a Gaussian distribution about these values.

Figure 7 shows example frames from one of the test sequences which contains six complete walking cycles. Clothing, viewing angle, and rate all differed from the training images. All cycles were correctly detected as shown in Figure 8. The graph on the right shows the expected phase $E[\phi|\mathbf{s}_t]$ while the graph on the left shows the probability that a cycle has completed which is defined to be

$$p(\mu^*) = \sum_{n=1}^{S} \begin{cases} \pi_t^{(n)} & \text{if } \mu \in \mathbf{s}_t^{(n)} \text{ and } \phi + 1 > \phi_{\max}, \\ 0 & \text{otherwise.} \end{cases}$$

Muybridge Experiment. We return now to the image sequence discussed in the Introduction. Figure 9 shows a sequence of images which covers slightly less than one walking cycle. The second row shows the estimated image motion within the predicted leg region using a conventional optical flow method [2]. The bottom row shows the expected flow field, $E[\vec{\mathbf{u}}(\vec{\mathbf{x}}; \mathbf{s}_t, t)|\mathbf{s}_t]$, generated from the spatio-temporal model. Note that the flow corresponding to each sample is constrained to be a valid walking motion.

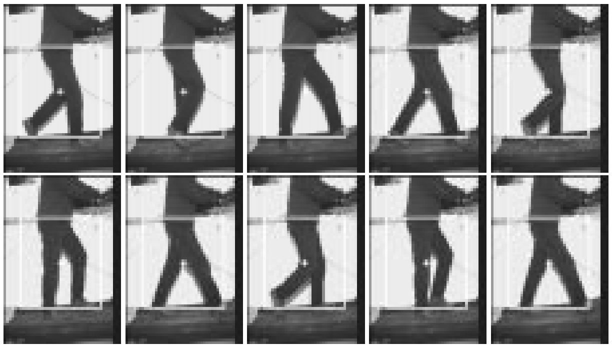

Figure 7: Test sequence, every 20 frames.

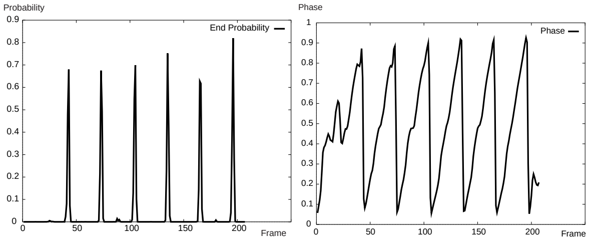

Figure 8: Recognition of walking cycles; left: probability that a cycle has completed; right: expected phase.

To better illustrate the behavior of the algorithm, Figure 10 shows the marginal distribution of the phase, ϕ, over 10 frames. Note that initially the distribution is uniform over phase. On the right in the figure is a plot of the mean of the coefficients, $E[a_i|\mathbf{s}_t]$. By the third frame the distribution is centered about the true phase and the mean trajectories of the coefficients approximate those of the model in Figure 6.

6.2 Mouth Motion

The next experiment introduces multiple motion models and allows the method to search over spatial location to automatically detect and track the mouth of a subject as they speak and move their head. The subject utters one of four test words (center, print, track, release). The spatial and temporal models are shown in Figures 3 and 4 respectively.

Evidence for the mouth position is constructed from a low-resolution average of the magnitude of the absolute temporal difference between frames. This is scaled to the size of the image and normalized. Peaks occur where there is motion but only those areas corresponding to mouth motions will have high likelihood. This simple scheme works well when there is limited head and camera motion though the assumption of conditional independence from the image derivatives is tenuous. Alternative cues such as color should be explored or importance sampling used [9].

The framework is applied to a 300-image sequence that was not part of the training set; the subject says "center print

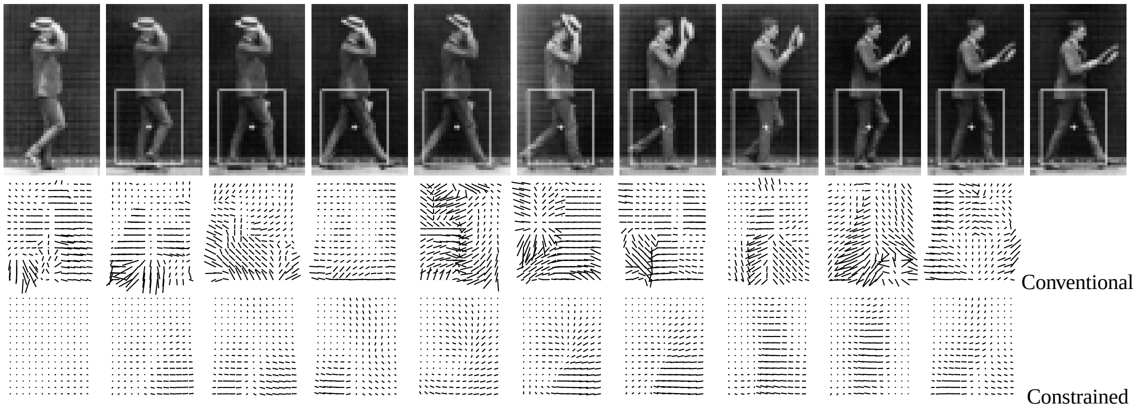

Figure 9: Muybridge image sequence; see text.

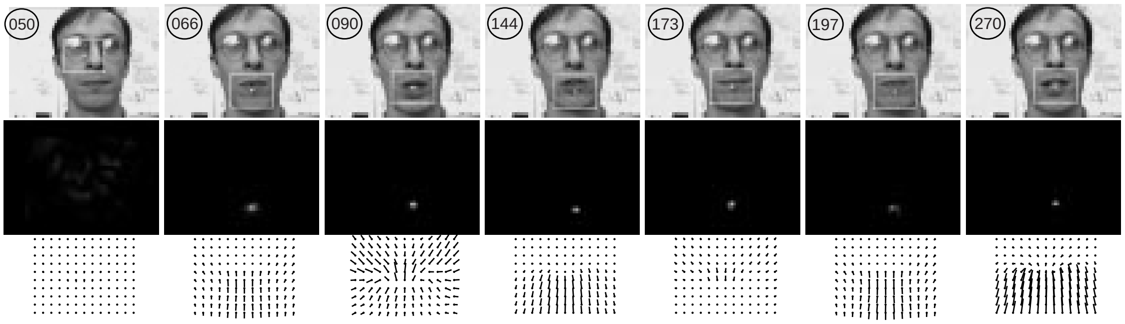

Figure 11: Top: expected mouth location. Middle: marginal distribution for position $\vec{\mathbf{p}}$. Bottom: mean optical flow.

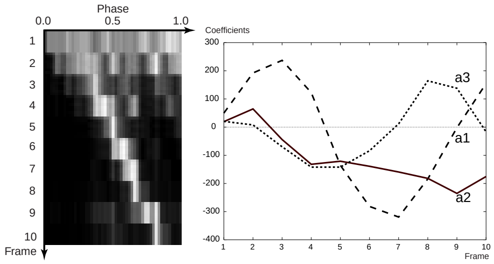

Figure 10: Muybridge. Left: marginal distribution for phase, ϕ, (brightness indicates probability). Right: mean of coefficients a_i.

track release center” followed by other utterances for which there was no temporal model. Sample images are shown in Figure 11 with a box superimposed on the expected mouth position $E[\vec{\mathbf{p}}|\mathbf{s}_t]$. Below each image is the marginal distribution of $p(s_t|\vec{\mathbf{z}}_t, m_t)$ shown for spatial position, $\vec{\mathbf{p}}$. When one of the spatio-temporal models fits the image motion, a peak is visible at the correct spatial position. Below this the mean flow for each of the mouth regions is shown.

Figure 12 (top) shows the marginal probability of each model as a function of time. Below that is the expected phase of each model. Note that sometimes the method quickly settles on a single model whereas in other cases (e.g. the utterance “track”), multiple hypotheses are maintained. Recognition is performed based on the probability that a model has terminated (bottom plot).

7 Conclusions

We have described parameterized spatio-temporal models for representing motion events in video sequences. We have proposed a Bayesian framework that permits the models to be non-linear or stochastic and a computational mechanism based on the Condensation algorithm for incrementally estimating a distribution over model parameters. The approach automatically detects and recognizes motion events based on image derivatives.

In the context of optical flow estimation, there has been a trend in the field moving from generic parameterized spatial models (affine) to object specific learned models. Here we continue that trend by adding object specific temporal models. Traditional estimation techniques are no longer applicable and this motivates the use of random sampling.

In the context of motion recognition, the method can be seen as providing a vocabulary of primitive optical flow events that can model fairly complex phenomena. The probabilistic formulation should allow this approach to be combined with higher-level recognition methods.

Note that we are asking a lot of our model: to detect and recognize events such as walking purely based on motion information. There are brightness cues that should be com-

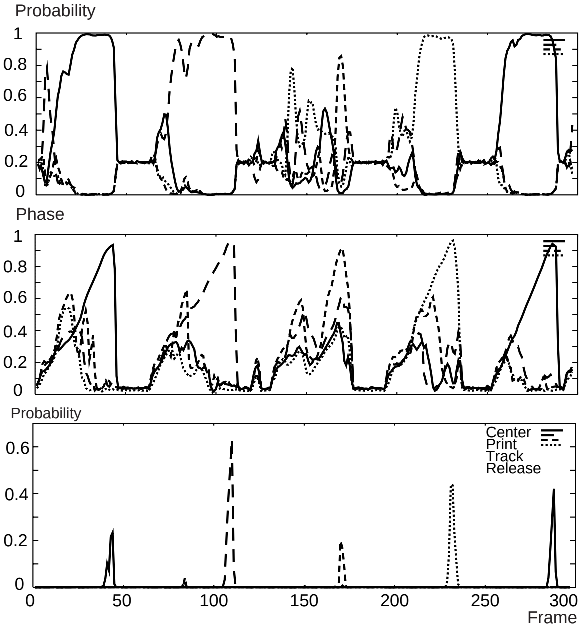

Figure 12: Mouth experiment; solid = "center," large dash = "print," small dash = "track," dot = "release." Top: marginal probability of each event. Middle: expected phase of each event. Bottom: probability of each event completing.

bined with motion to further constrain the problem. This is necessary since, the number of samples required to represent the distribution may grow exponentially with the size of the parameter space. With 35000 samples, the computation takes approximately one minute per frame.

Immediate topics for future work include expanding the experimentation to include more models of activities (e.g. walking from various view points, sitting, running, etc.), multi-part and multi-scale models (e.g. a low resolution person model combined with a high resolution leg model), adding models of image appearance change in addition to motion, and adding a search over small rotations. Additionally, we are developing stochastic models of motion texture which fit naturally in the framework described here. Temporal stochastic models based on HMM's may also be exploited.

The models and mechanisms described here shift the focus of the optical flow problem: movement in the image sequence should be "explained" using available models of the world. This motion explanation problem focuses on characterizing image brightness variation rather than accurately estimating image motion.

Acknowledgements. This paper builds on previous work with A. Jepson, Y. Yacoob, and D. Fleet and their contributions are gratefully acknowledged. The author thanks L. Bretzner and W. Rucklidge who provided comments that helped clarify the paper, H. Christensen and J. Eklundh who provided a wonderful research environment at KTH, and Y. Yacoob who provided the treadmill image sequences.

References

[1] A. Bab-Hadiashar and D. Suter. Optic flow calculation using robust statistics. *CVPR*, pp. 988–993, 1997.

[2] M. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *CVIU*, 63(1):75–104, Jan. 1996.

[3] M. Black and A. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. *ECCV*, LNCS vol. 1406, pp. 909–924, 1998.

[4] M. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *IJCV*, 25(1):23–48, 1997.

[5] M. Black, Y. Yacoob, A. Jepson, and D. Fleet. Learning parameterized models of image motion. *CVPR*, pp. 561–567, 1997.

[6] C. Bregler and J. Malik. Tracking people with twists and exponential maps. *CVPR*, 1998.

[7] J. Davis and A. Bobick. The representation and recognition of action using temporal templates. *CVPR*, pp. 928–934, 1997.

[8] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *ECCV*, LNCS vol. 1064, pp. 343–356, 1996.

[9] M. Isard and A. Blake. ICondensation: Unifying low-level and high-level tracking in a stochastic framework. *ECCV*, LNCS vol. 1406, pp. 893–908, 1998.

[10] S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. *Int. Conf. Automatic Face and Gesture Recog.*, pp. 38–44, 1996.

[11] J. Little and J. Boyd. Recognizing people by their gait: The shape of motion. *Videre: J. of Comp. Vis. Res.*, 1(2):1–32, 1998.

[12] F. Liu and R. W. Picard. Finding periodicity in space and time. *ICCV*, pp. 374–381, 1998.

[13] E. Muybridge. *The Human Figure in Motion*. Dover Pub., Inc., NY, 1955.

[14] S. Niyogi and E. Adelson. Analyzing and recognizing walking figures in xyt. *CVPR*, pp. 469–474, 1994.

[15] R. Polana and R. C. Nelson. Detecting activities. *CVPR*, pp. 2-7, 1993.

[16] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. *ICCV*, pp. 374–381, 1998.

[17] Y. Yacoob and L. Davis. Learned temporal models of image motion. *ICCV*, pp. 446–453, 1998.

[18] Y. Yacoob and L. Davis. Temporal multi-scale models for flow and acceleration. *CVPR*, pp. 921–927, 1997.

[19] A. Yuille, P-Y. Burgi, and N. Grzywacz. Visual motion estimation and prediction: A probabilistic network model for temporal coherence. *ICCV*, pp. 973–978, 1998.