CAMERA MOTION ESTIMATION USING PARTICLE FILTERS

Symeon Nikitidis, Stefanos Zafeiriou and Ioannis Pitas *
Aristotle University of Thessaloniki, Department of Informatics, Box 451, 54124 Thessaloniki, Greece
{*nikitidis,dralbert,pitas*}*@aiia.csd.auth.gr*

November 22, 2007

Abstract

In this paper a novel algorithm for estimating the parametric form of the camera motion is proposed. A novel stochastic vector field model is presented which can handle smooth motion patterns derived from long periods of stable camera movement and also can cope with rapid motion changes and periods where camera remains still. A set of rules for robust and online updating of the model parameters is also proposed, based on the Expectation Maximization algorithm. Finally, we fit this model in a particle filters framework, in order to predict the future camera motion based on current and prior knowledge.

Camera Motion Estimation, Vector Field Model, Particle Filtering, Expectation Maximization Algorithm.

1 Introduction

Video, in contrary with image, possesses valuable information since it extends spatial information and records the evolution of events over time. This dynamic property, has been extensively investigated by the scientific community for semantic characterization and discrimination of videos streams. In particular, considerable interest has been focused in extracting motion related information such as object and camera motion. Moving objects trajectories have been used for video retrieval in [Hu et al., 2007] as well as, camera motion pattern characterization has been efficiently applied for video data indexing and retrieval in [Tan et al., 2000, Kim et al., 2000]. In [Duan et al., 2006], the motion vectors field is used as a camera motion representation and the detected motion pattern is classified using Support Vector Machines (SVMs) in one of the following classes: zoom, pan, tilt and rotation. In [Tan et al., 2000] and [Kim et al., 2000] camera motion estimation within video shots is performed in the compressed MPEG video streams, without full frame decompression, using the motion vector fields acquired from the P- and B- video frames. These methods rely on the exploitation of motion vectors distribution or on a few representative global motion parameters. However, one of the main shortcomings of these approaches is that they are generally not resilient in the presence of mobile objects of significant size and data outliers.

Camera motion can be assumed as a dynamic system, whose state $\boldsymbol{\theta}_t$ is described at time t by the state vector: $\boldsymbol{\theta}_t = \begin{bmatrix} m_1 & m_2 & m_3 & m_4 & m_5 & m_6 \end{bmatrix}^T$ where parameters $\{m_1, m_2, m_3, m_4, m_5, m_6\}$ correspond to the affine transform coefficients, containing

[*] Research was supported by the project SHARE: Mobile Support for Rescue Forces, Integrating Multiple Modes of Interaction, EU FP6 Information Society Technologies, Contract Number FP6-004218.

all the relevant information required to describe the camera motion within frames. Our goal is to recursively estimate the system state $\boldsymbol{\theta}_t$ from noisy measurements $\mathbf{Y}_t$, obtained by an observation model. To tackle this problem, we propose a novel stochastic vector field model applied in a particle filters framework.

2 Problem Formulation

In order to measure the displacement between two consecutive frames, we employ the motion vectors derived by a motion compensation algorithm such as block matching [Jain and Jain, 1981]. A motion vector $\mathbf{v}^i = [v_x^i \quad v_y^i]^T$ represents the displacement of the i-th block in relative coordinates between two consecutive video frames f_{t-1} and f_t as: $x'_i = x_i + v_x^i,\ \ y'_i = y_i + v_y^i$ where (x_i, y_i) and (x'_i, y'_i) are the coordinates of i-th block center at frame f_{t-1} and f_t, respectively.

We can represent the displacement of the i_{th} block by a 2D affine transform as:

$$\begin{bmatrix} x'_i \\ y'_i \\ 1 \end{bmatrix} = \begin{bmatrix} m_1 & m_2 & m_3 \\ m_4 & m_5 & m_6 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} v_x^i \\ v_y^i \\ 0 \end{bmatrix} = \begin{bmatrix} m_1 - 1 & m_2 & m_3 \\ m_4 & m_5 - 1 & m_6 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix}. \tag{1}$$

We seek an affine transformation matrix $\mathbf{M}$ such that to approximate $\mathbf{BM} \approx \mathbf{B} + \mathbf{V}$, where $\mathbf{B}$ is a $n \times 3$ matrix (n is the number of blocks that each frame has been divided to) containing the center coordinates of each block. Matrix $\mathbf{V} = [\mathbf{v}_x\ \mathbf{v}_y\ \mathbf{0}]$ contains the motion vectors, where $\mathbf{v}_x = [v_x^1 \quad v_x^2 \ldots v_x^n]^T$ and $\mathbf{v}_y = [v_y^1 \quad v_y^2 \ldots v_y^n]^T$ are $n \times 1$ vectors containing the motion vectors residuals along to the x and y axes, respectively. $\mathbf{M} = [\mathbf{M}_x\ \mathbf{M}_y\ \mathbf{e}]$ is the 3×3 affine transformation matrix, where $\mathbf{M}_x = [m_1 \quad m_2 \quad m_3]^T$, $\mathbf{M}_y = [m_4 \quad m_5 \quad m_6]^T$ and $\mathbf{e} = [0 \quad 0 \quad 1]^T$.

3 Online Vector Field Model

The presented *Online Vector Field Model* (*OVFM*) exploits temporal characteristics of camera motion. $OVFM$ is time-varying and comprises of three different components $OVFM_t = \{\mathbb{S}_t, \mathbb{W}_t, \mathbb{L}_t\}$ which are combined in a probabilistic mixture model.

3.1 Probabilistic Mixture Model

The stable component $\mathbb{S}_t = \{\mathbf{S}_{t,x}, \mathbf{S}_{t,y}\}$ learns a smooth camera motion pattern obtained from a relatively long period of the video sequence. The component $\mathbb{S}_t$ comprises of vectors $\mathbf{S}_{t,x} = [s_{t,x}^1 \quad s_{t,x}^2 \ldots s_{t,x}^n]^T$ and $\mathbf{S}_{t,y} = [s_{t,y}^1 \quad s_{t,y}^2 \ldots s_{t,y}^n]^T$ where values $s_{t,x}^j$ and $s_{t,y}^j$ contain the block j displacement momentum along the x and y axes, respectively. The wander component $\mathbb{W}_t = \{\mathbf{W}_{t,x}, \mathbf{W}_{t,y}\}$, identifies sudden motion changes, and adapts with a short time observation sequence, as a two frame motion change model. Vectors $\mathbf{W}_{t,x} = [w_{t,x}^1 \quad w_{t,x}^2 \ldots w_{t,x}^n]^T$ and $\mathbf{W}_{t,y} = [w_{t,y}^1 \quad w_{t,y}^2 \ldots w_{t,y}^n]^T$ contain the motion vectors residuals, along the x and y axes, respectively. Finally, the lost component $\mathbb{L}_t = \{\mathbf{L}_{t,x}, \mathbf{L}_{t,y}\}$ is fixed and represents the ideal stationary video scene.

We model the probability density function for the $\mathbb{S}_t$, $\mathbb{W}_t$ and $\mathbb{L}_t$ components with the bivariate Gaussian distribution $N(\mathbf{v}^j; \boldsymbol{\mu}_{c,t}^j, \boldsymbol{\Sigma}_{c,t}^j)\ c \in \{\mathbb{S}_t, \mathbb{W}_t, \mathbb{L}_t\}$, where $\boldsymbol{\Sigma}_{c,t}^j$ is a 2×2 covariance matrix referred to c-th component j-th motion vector containing the two random variables v_x^j and v_y^j, $\boldsymbol{\mu}_{c,t}^j$ denotes the mean value of the j-th motion vector. $OVFM_t$ combines probabilistically components $\mathbb{S}_t$, $\mathbb{W}_t$ and $\mathbb{L}_t$ according to the formula:

$$P(\mathbf{Y}_t|\boldsymbol{\theta}_t) = \prod_{j=1}^{n} \Big\{ P(\mathbf{v}_t^j|\mathbb{S}_t^j) + P(\mathbf{v}_t^j|\mathbb{W}_t^j) + P(\mathbf{v}_t^j|\mathbb{L}_t^j) \Big\} = \prod_{j=1}^{n} \Big\{ \sum_{c=\mathbb{S},\mathbb{W},\mathbb{L}} m_{c,t,xy}^j N\Big(\mathbf{v}_t^j; \boldsymbol{\mu}_{c,t}^j, \boldsymbol{\Sigma}_{c,t}^j\Big) \Big\}, \tag{2}$$

where $\mathbf{Y}_t = [\mathbf{v}_t^1 \ldots \mathbf{v}_t^n]^T$ is the observation data derived for state $\boldsymbol{\theta}_t$. The mixing probabilities $m_{c,t,xy}^j$ regulate the contribution that each component j-th motion vector makes to the complete observation likelihood at time t and n is the number of motion vectors. $OVFM_t$ is embedded in the particle filters framework evaluating each potential future state of the system. A state estimate $\hat{\boldsymbol{\theta}}_t^i$ is generated by first drawing a Gaussian noise sample U_{t-1}^i and applying the state transition function $\hat{\boldsymbol{\theta}}_t^i = E_{t-1}(\boldsymbol{\theta}_{t-1}^i, U_{t-1}^i)$. Each state estimate $\hat{\boldsymbol{\theta}}_t^i$ determined by particle i is being evaluated with respect to the available motion representation in $OVFM_t$, by computing the observation likelihood according to equation (2). Weights are assigned to particles by applying the Sequential Importance Re-sampling filter (SIR) proposed in [Gordon et al., 1993], as: $w_t^i \propto P(\mathbf{Y}_t^i|\boldsymbol{\theta}_t^i)$, $\quad w_t^i = \frac{w_t^i}{\sum_{i=1}^N w_t^i}$.

3.2 Online Model Update

We assume that $OVFM_t$ has limited memory over the past motion observations and when newer information is available, previous knowledge is forgotten and is combined with newer observations using the exponential envelop $E_t(k) = \alpha e^{(-(t-k)/\tau)}$ where $\tau = n_s/\log 2$, n_s is the envelope's half life in video frames and parameter α is defined as $\alpha = 1 - e^{-1/\tau}$ in order the ownership posterior probabilities and the mixing probabilities to sum to 1. The posterior ownership probabilities $O_{c,t}$ denote the contribution that each component motion vector makes to the complete observation likelihood. Ownerships are evaluated by applying the EM algorithm in [Dempster et al., 1977] as:

$$\begin{aligned} O_{c,t,xy}^j &\propto m_{c,t,xy}^j N(\mathbf{v}_t^j; \boldsymbol{\mu}_{c,t}^j, \boldsymbol{\Sigma}_{c,t}^j), \qquad O_{c,t,x}^j \propto m_{c,t,x}^j \mathcal{N}\left(v_{t,x}^j; \mu_{c,t,x}^j, (\sigma_{c,t,x}^j)^2\right) \\ O_{c,t,y}^j &\propto m_{c,t,y}^j \mathcal{N}\left(v_{t,y}^j; \mu_{c,t,y}^j, (\sigma_{c,t,y}^j)^2\right) \end{aligned} \tag{3}$$

where $\mathcal{N}\left(v_{t,x}^j; \mu_{c,t,x}^j, (\sigma_{c,t,x}^j)^2\right)$ is the normal density function. The ownerships are subsequently used for updating the mixing probabilities as:

$$\begin{aligned} m_{c,t+1,x}^j &= \alpha O_{c,t,x}^j + (1-\alpha) m_{c,t,x}^j, \qquad m_{c,t+1,y}^j = \alpha O_{c,t,y}^j + (1-\alpha) m_{c,t,y}^j \\ m_{c,t+1,xy}^j &= \alpha O_{c,t,xy}^j + (1-\alpha) m_{c,t,xy}^j. \end{aligned} \tag{4}$$

We compute the new mean values and the new covariance matrices for each motion vector by utilizing the first and second order data moments computed as:

$$\begin{aligned} M_{1,t+1,x}^j &= \alpha O_{\mathbb{S},t,x}^j v_{t,x}^j + (1-\alpha) M_{1,t,x}^j, \qquad M_{1,t+1,y}^j = \alpha O_{\mathbb{S},t,y}^j v_{t,y}^j + (1-\alpha) M_{1,t,y}^j \\ M_{1,t+1,xy}^j &= \alpha O_{\mathbb{S},t,xy}^j v_{t,x}^j v_{t,y}^j + (1-\alpha) M_{1,t,xy}^j \\ M_{2,t+1,x}^j &= \alpha O_{\mathbb{S},t,x}^j (v_{t,x}^j)^2 + (1-\alpha) M_{2,t,x}^j, \quad M_{2,t+1,y}^j = \alpha O_{\mathbb{S},t,y}^j (v_{t,y}^j)^2 + (1-\alpha) M_{2,t,y}^j. \end{aligned} \tag{5}$$

The stable component is updated using the first order data moments as:

$$s_{t+1,x}^j = \mu_{\mathbb{S},t+1,x}^j = \frac{M_{1,t+1,x}^j}{m_{\mathbb{S},t+1,x}^j}, \quad s_{t+1,y}^j = \mu_{\mathbb{S},t+1,y}^j = \frac{M_{1,t+1,y}^j}{m_{\mathbb{S},t+1,y}^j}. \tag{6}$$

The stable component new covariance matrices are evaluated as:

$$\begin{aligned} (\sigma_{\mathbb{S},t+1,x}^j)^2 &= \frac{M_{2,t+1,x}^j}{m_{\mathbb{S},t+1,x}^j} - (s_{t+1,x}^j)^2, \quad (\sigma_{\mathbb{S},t+1,y}^j)^2 = \frac{M_{2,t+1,y}^j}{m_{\mathbb{S},t+1,y}^j} - (s_{t+1,y}^j)^2 \\ (\sigma_{\mathbb{S},t+1,xy}^j)^2 &= \frac{M_{1,t+1,xy}^j}{m_{\mathbb{S},t+1,xy}^j} - (s_{t+1,x}^j)(s_{t+1,y}^j). \end{aligned} \tag{7}$$

The wander component contains the current motion vectors, since it adapts as a two frame motion change model, while covariance matrices for the wander and lost components are updated according to stable component's covariance matrices in order to avoid some prior preference in either component.

4 Experimental results

We have evaluated the efficiency of our algorithm through experimental testing. Experiments have been conducted in a dataset comprised of infrared video streams captured by a hand-held video camera. We present the results obtained by applying our method in a video sequence containing 485 frames, where camera performs a 360 degrees spin. Figure 1 presents the variation of the estimated affine coefficients describing the translation over the x axis (dotted black line) and y axis (solid gray line) as the video stream evolves, while at specific moments the respective video frames are provided for visual confirmation of the obtained results. As it is depicted both parameters balances around zero during the first 25 frames since the camera remains almost still. A radical incensement in the coefficient describing translation over the x axis occurs from frame 26 and until the end of the video stream since camera starts to spin. On the other hand, the coefficient that corresponds to translation over the y axis continuously balances around zero since there is minimum movement towards that direction.

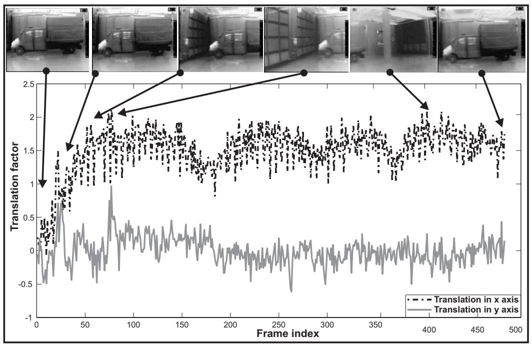

Figure 1: Variation of the translation affine coefficients.

5 Conclusion

In this paper a novel camera motion estimation method based in motion vector fields exploitation is proposed. The features that distinguish our method from other proposed camera motion estimation techniques are: 1) the integration of a novel stochastic vector field model, 2) the incorporation of the vector field model inside a particle filters framework enabling the method to estimate the future camera movement.

References

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[Duan et al., 2006] Duan, L.-Y., Jin, J. S., Tian, Q., and Xu, C.-S. (2006). Nonparametric motion characterization for robust classification of camera motion patterns. *IEEE Transactions on Multimedia*, 8(2):323–340.

[Gordon et al., 1993] Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian bayesian state estimation. In *Radar and Signal Processing, IEEE Proceedings F*, volume 140, pages 107–113.

[Hu et al., 2007] Hu, W., Xie, D., Fu, Z., Zeng, W., and Maybank, S. (2007). Semantic-based surveillance video retrieval. *IEEE Transactions on Image Processing*, 16(4):1168 – 1181.

[Jain and Jain, 1981] Jain, J. R. and Jain, A. K. (1981). Displacement measurement and its application in interframe image coding. *IEEE Transactions on Communications*, COM-29(12):1799–1808.

[Kim et al., 2000] Kim, J.-G., Chang, H.-S., Kim, J., and Kim, H.-M. (30 July-2 Aug. 2000). Efficient camera motion characterization for MPEG video indexing. In *ICME '00*, volume 2, pages 1171 – 1174.

[Tan et al., 2000] Tan, Y.-P., Saur, D. D., Kulkarni, S. R., and Ramadge, P. J. (2000). Rapid estimation of camera motion from compressed video with application to video annotation. *IEEE Transactions on Circuits and Systems for Video Technology*, 10(1):133–145.