TEMPORAL REGISTRATION OF VIDEO SEQUENCES [1]

Hui Cheng

Sarnoff Corporation
hcheng@sarnoff.com

ABSTRACT

In this paper, we propose a temporal registration algorithm for video sequences. This algorithm is developed based on a frame-level model of the temporal misalignment often introduced by video processing algorithms, such as compression, frame rate conversion or by video capturing. With this model, the temporal registration is formulated as a constrained minimization of a matching cost, and it is solved using dynamic programming. The proposed temporal registration can also be extended to integrate both spatial registration and histogram registration. In addition, prior information about the application can be easily incorporated in the form of contextual cost. Experimental results on both synthetic and real processed video sequences with temporal misalignments have shown that the proposed algorithm is effective and robust to a wide range of noises.

1. INTRODUCTION

For many applications, such as watermark detection [1] and reference-based video quality measurement [2], a processed video sequence needs to be registered to the original sequence. For example, to detect watermarks embedded in pirated videos that are shot using camcorder, we may need to register the captured video to the original one displayed in the theater. Another area where we need video registration is reference-based video quality measurement. To ensure quality of service (QoS), it is often necessary to measure the quality degradation between the original video and the one received by a client. The received video is often a processed version of the original video. Therefore, to achieve a meaningful reference-based quality measurement, the received video needs to be first registered to the original video sequence.

The misalignment between a processed video and the original one is generally a combination of spatial misalignment, temporal misalignment and histogram misalignment. Spatial misalignment is the result of spatial manipulation of a video sequence, such as warping, cropping and resizing. The main causes of temporal misalignment are the change of temporal resolution, such as frame rate conversion (e.q. 3-2 pull down), and frame dropping or frame repeat used by video compression algorithms (i.e. MPEG-4). The video capturing process also causes temporal misalignment, because the displaying and the capturing generally are not synchronized and operate at different frame rates. In addition, processed videos in general have different color histograms from the original videos. This is often the result of video processing, such as compression, filtering or gamma changes. It can also be the result of white balance or automatic gain control (AGC) in camcorder capture.

To correct the three types of misalignment, spatial, temporal and histogram registration are needed. Spatial and histogram registrations have been studied by many researchers. However, few studies have been done on temporal registration. Lu [3] proposed a temporal registration for video quality measurement. It can recover global offset between two videos. The global offset is estimated by maximizing the normalized correlation between a temporal activity signatures extracted from each sequence. In [4], Capsi and Irani uses direct search for recovering sequence-level temporal misalignment, such as fixed shift or fixed frame rate conversion.

In this paper, we propose a temporal registration algorithm for videos. The proposed algorithm formulates the temporal registration as a frame-level constrained minimization of a matching cost and solves it using dynamic programming [5]. It is similar to the algorithm proposed for word segmentation in sentences in speech recognition [6].

This temporal registration algorithm can recover temporal misalignment at frame-level instead of at sequence level. Therefore, it can recover a much wider range of temporal misalignment, such as frame drop or repeat. One advantage of the proposed framework is that it can be generalized to incorporate both spatial and histogram registration. In addition, it not only allows the registration to be performed according to the video data, but also allows the integration of prior knowledge of what the registration should be in the form of contextual cost. Therefore, further improvement of both the accuracy and the robustness are possible. The contextual cost can also be easily adjusted according to the application by using domain specific contextual information.

2. MODEL FOR TEMPORAL VIDEO PROCESSING

In order to develop an effective temporal registration algorithm, we first model the temporal processing aspect incorporated in most of the temporally misaligned video sequences. We denote the original video frames (When input video is not progressive, use fields instead of frames.) as I_i and processed video frames as J_j. Then, most of the temporal misalignment that need to be registered can be modeled using a simple 2-frame integration model. That is,

$$\begin{aligned} J_j &\equiv \varphi\left(I_{\alpha(j)}, I_{\alpha(j)-1}; \lambda_{j,\alpha(j)}\right) \\ &= \lambda_{j,\alpha(j)} \cdot I_{\alpha(j)} + \left(1-\lambda_{j,\alpha(j)}\right) \cdot I_{\alpha(j)-1}, \end{aligned} \tag{1}$$

[1] This work was performed under the support of the U.S. Department of Commerce, National Institute of Standards and Technology, Advanced Technology Program, Cooperative Agreement Number 70NANB1H3036.

where $\alpha(j)$ is the matching index that maps a processed video frame index to a frame index in the original video, and $\lambda_{j,\alpha(j)}$ is the weight of $I_{\alpha(j)}$ in the frame integration. $\lambda_{j,\alpha(j)}$ is larger than 0 and smaller or equal to 1, and it can be estimated efficiently using a closed-form formula given in section 3.1.

Although the 2-frame integration model is simple, it represents many widely used frame-level temporal operations, such as frame drop / frame repeat used in video compression (i.e. MPEG-4) or frame rate conversion (e.q. 3-2 pull-down). It is also a good model for the video capture process. Because the displaying and the capturing are not synchronized, most captured frames are a linear combination of two consecutive displayed frames as shown in Figure 1. In this case, λ_{ji} is the percentage of exposure of I_i during the capture of J_j.

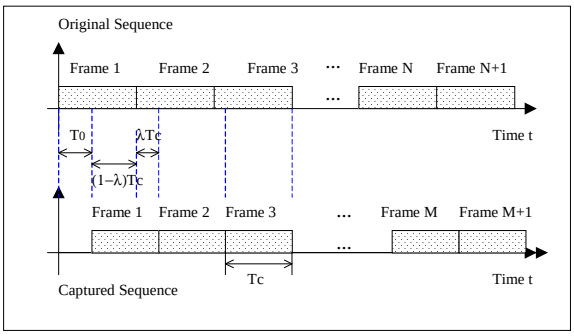

Figure 1. Frame integration during video capturing.

In Figure 2, we show an example of representing a processed video using this temporal processing model. Based on the values of $\alpha(j)$ and $\lambda_{j,\alpha(j)}$ in Figure 2, the captured frames 0 and 1 are the original frames 1 and 3, respectively. The captured frame 2 is a repeat of captured frame 1. Frame 3 of the capture sequence is a frame integration of frames 6 and 7 in the original sequence. Since no captured frames are mapped to original frames 0, 2, 4 and 5, they are dropped frames.

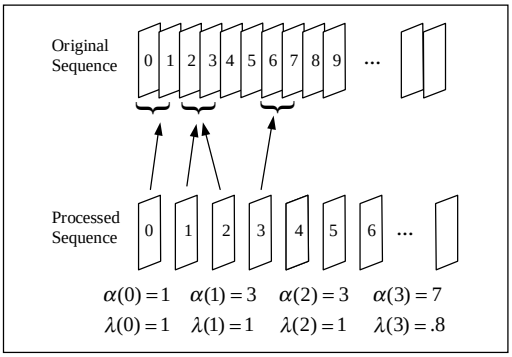

Figure 2. An example of representing a processed video using the proposed temporal processing model.

3. TEMPORAL VIDEO REGISTRATION

With the above model, the temporal registration can be formulated as a minimization of the matching cost. That is, given the original and the processed videos, estimate the matching indices, $\alpha(j)$ and the matching parameters $\lambda_{j,\alpha(j)}$ that minimize the distortion between J_j and the model prediction of J_j from the original video over a window and all possible combinations of $\lambda_{j,\alpha(j)}$ and $\alpha(j)$. However, the minimization is subject to a causal constraint on $\alpha(j)$. That is, for any given j_1 and j_2, if $j_1 < j_2$, then $\alpha(j_1) \le \alpha(j_2)$. The causal constraint requires the same temporal ordering among the processed video as the original one. This is enforced by most video processing algorithms. In the case of video capture, it means no frames displayed in the past can be captured in the future. In addition, in this paper, we adopt the mean squared error (MSE) as the distortion measure. Therefore, the registration parameters $\left[\lambda^*(j), \alpha^*(j)\right]$ are computed as

$$\left[\lambda^*(0), \alpha^*(0), \cdots, \lambda^*(M), \alpha^*(M)\right] = \underset{\substack{[\lambda(0),\cdots,\lambda(M),\\ \alpha(0)\le\cdots\le\alpha(M)]}}{\arg\min} \sum_{j=0}^{M} \left\| J_j - \varphi\left[I_{\alpha(j)-1}, I_{\alpha(j)}; \lambda_j\right]\right\|^2, \quad (2)$$

where M is the number of captured frames, and $\lambda(j)$ is $\lambda_{j,\alpha(j)}$. Since $\lambda(j)$ for $j = 0, \cdots, M$ can be optimized independently of each other, and there is only causal dependency among $\alpha(j)$, the optimization defined in (2) can be solved using dynamic programming [5].

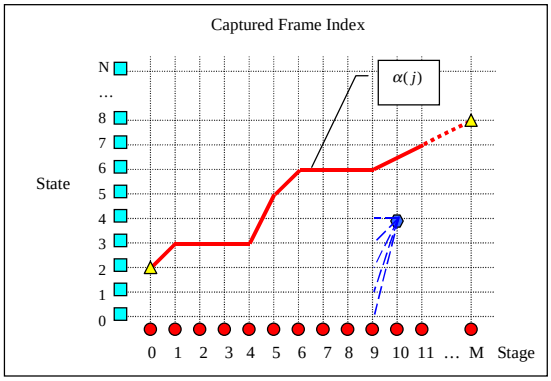

Figure 3. Temporal registration using dynamic programming. Circles are captured frames. Squares are original frames. A feasible path is marked by a solid line. All feasible paths from the point marked by the hexagon are shown using dash lines.

To solve (2) using dynamic programming, we first partition the minimization into *stages* according to the index of the processed frame, j. The *state* for each stage is the original frame index, denoted as i. As shown in Figure 3, in a grid defined by stages and states, $\alpha(j)$ defines a mapping from stages to states. We call this mapping a path from one stage to another. Therefore, the solution of (2) is a path from stage 0 to stage M that has the minimal accumulated mean squared error.

However, because of the causal constraint, $\alpha(0) \le \alpha(1) \le \cdots \le \alpha(M)$, the solution to (2) can only be a

feasible path, i.e. paths that are monotonically non-decreasing in value. In Figure 3, we show a feasible path from stage 0 to stage M using a solid line. We also show all feasible paths that pass a grid point marked by the hexagon using dash lines. Therefore, the solution to (2) is a monotonically non-decreasing path from stage 0 to stage M that has the minimal accumulated mean squared error (MSE).

Denote the accumulated MSE over a feasible path from stage 0 to stage M as $\delta(M)$. Then,

$$\begin{aligned}\delta(M) &\equiv \min_{\substack{[\lambda(0),\cdots,\lambda(M),\\ \alpha(0)\le\cdots\le\alpha(M)]}} \sum_{j=0}^{M} \left\| J_j - \varphi\left[I_{\alpha(j)-1}, I_{\alpha(j)}; \lambda(j)\right]\right\|^2 \\ &= \min_{\alpha(M-1)\le\alpha(M)} \left\{ \min_{\substack{[\lambda(0),\cdots,\lambda(M-1),\\ \alpha(0)\le\cdots\le\alpha(M-1)]}} \sum_{j=0}^{M-1} \left\| J_j - \varphi\left[I_{\alpha(j)-1}, I_{\alpha(j)}; \lambda(j)\right]\right\|^2 \right. \\ &\quad \left. + \min_{\lambda(M)} \left\| J_M - \varphi\left[I_{\alpha(M)-1}, I_{\alpha(M)}; \lambda(M)\right]\right\|^2 \right\} \\ &= \min_{\alpha(M-1)\le\alpha(M)} \left\{ \delta(M-1) + \min_{\lambda(M)} \left\| J_M - \varphi\left[I_{\alpha(M)-1}, I_{\alpha(M)}; \lambda(M)\right]\right\|^2 \right\}\end{aligned} \tag{3}$$

Therefore, the dynamic programming contains the following three steps:

(1) Local minimization at each node (j,i).

$$\lambda\left(j \mid \alpha(j)=i\right) = \arg\min_{\lambda(j)} \left\| J_j - \varphi\left[I_{i-1}, I_i; \lambda_{ji}\right]\right\|^2 \tag{4}$$

(2) Recursively, as shown in Eq. (3), compute $\delta(j)$, for $j = 0,1,...,M$.

(3) After $\delta(M)$, the minimal accumulated MSE for the last stage is calculated, back trace to get $\left[\lambda^*(0), \alpha^*(0), \cdots, \lambda^*(M), \alpha^*(M)\right]$.

3.1 Minimization of Local Prediction Error

As shown in the first step of the above dynamic programming algorithm, we need to minimize the local prediction error defined in Eq (4). Substituting Eq (1) into Eq (4), we have

$$\begin{aligned}\sigma(i,j) &\equiv \min_{\lambda_{ji}} \left\| J_j - \varphi[I_i, I_{i-1}; \lambda_{ji}]\right\|^2 \\ &= \min_{\lambda_{ji}} \left\| J_j - \left(\lambda_{ji} I_k + (1+\lambda_{ji}) I_{k-1}\right)\right\|^2\end{aligned} \tag{5}$$

for any given captured frame J_j and original frames I_i and I_{i-1}. Let $\varepsilon(F,G)$ be the mean squared error between two images, F and G, of size $M \times N$.

$$\varepsilon(F,G) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left(F_{m,n} - G_{m,n}\right)^2$$

For three images, F, H and G, we define $\varsigma(F;H,G)$ as the "cross correlation" between the differences $(F_{m,n} - G_{m,n})$ and $(F_{m,n} - H_{m,n})$. That is,

$$\varsigma(F;G,H) = \frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \left(F_{m,n} - G_{m,n}\right)\left(F_{m,n} - H_{m,n}\right).$$

Then, the λ_{ji} minimizes Eq (5), λ^*_{ji}, is

$$\lambda^*_{ji} = \max\left(0, \min\left(1, \frac{\varepsilon(J_j, I_i) - \varsigma(J_j; I_i, I_{i-1})}{\varepsilon(J_j, I_i) - 2\varsigma(J_j; I_i, I_{i-1}) + \varepsilon(J_j, I_{i-2})}\right)\right)$$

and the minimal MSE is

$$\begin{aligned}&\min_{0\le\lambda_{ji}\le 1} \left\| J_j - \varphi\left(I_i, I_{i-1}; \lambda_{ji}\right)\right\|^2 \\ &= \min\left(\frac{\varepsilon(J_j, I_i)\varepsilon(J_j, I_{i-1}) - \varsigma^2(J_j; I_i, I_{i-1})}{\varepsilon(J_j, I_i) - 2\varsigma(J_j; I_i, I_{i-1}) + \varepsilon(J_j, I_{i-2})}, \right. \\ &\qquad \left. \varepsilon(J_j, I_i), \varepsilon(J_j, I_{i-1})\right)\end{aligned}$$

3.2 Contextual Constraints

Video registration is an ill-posed inverse problem. For a given original video and a processed video, there may exist more than one solution. However, due to the nature of the prior knowledge of the application, the solutions to the same problem may have significantly different probabilities. For example, frame repeat or frame drop is usually used infrequently, and they are seldom repeated more than once. When there are a large number of similar frames, they are more likely from a scene of little motion than caused by consecutive uses of frame repeat. Therefore, contextual constraints, the prior knowledge of what a solution must satisfy, can be used of to reduce the solution space, improve the accuracy and increase the robustness against noises. Using contextual constraints in the form of cost functions, $C\left(\lambda(0),\alpha(0),\cdots,\lambda(M),\alpha(M)\right)$, the optimization problem in (2) can be extended to

$$\begin{aligned}&\left[\lambda^*(0),\alpha^*(0),\cdots,\lambda^*(M),\alpha^*(M)\right] \\ &= \arg\min_{\substack{[\lambda(0),\cdots,\lambda(M),\\ \alpha(0)\le\cdots\le\alpha(M)]}} \sum_{j=0}^{M} \left\| J_j - \varphi\left[I_{\alpha(j)-1}, I_{\alpha(j)}; \lambda(j)\right]\right\|^2 \\ &\qquad + C\left(\lambda(0),\alpha(0),\cdots,\lambda(M),\alpha(M)\right)\end{aligned} \tag{10}$$

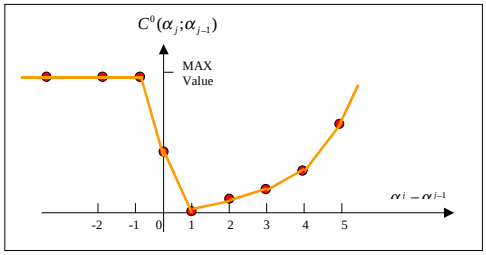

Figure 4. Plot of contextual cost, $C\left(\lambda(j),\alpha_j;\lambda(j-1),\alpha_{j-1}\right)$.

One contextual constraint already used is the causal constraints on matching indices. The causal constraint enforces the non-decreasing temporal ordering in the matching indices. However, not all monotonically non-decreasing paths among the state space are feasible. For example, when a change of frame occurs during the capture, $0<\lambda<1$, if $\alpha_j = \alpha_{j+1}$, then either $\lambda_j = \lambda_{j+1}$ (frame repeat), or $0 < \lambda_j < \lambda_{j+1} = 1$. That is, if two consecutive captured frames correspond to the same set of original frames, then except for frame repeat, the integration of

the original frames can only happen during the capture of the first frame. Other results, such as $0 < \lambda_j < \lambda_{j+1} < 1$ or $\lambda_j > \lambda_{j+1}$ are invalid. One contextual cost function is shown in Figure 4. It enforces the following contextual constraints:

(1) $C(\lambda(j),\alpha_j;\lambda(j-1),\alpha_{j-1})$ is set to maximal distortion when $\alpha_j < \alpha_{j-1}$, to enforces the causal constraint.

(2) $C(\lambda(j),\alpha_j;\lambda(j-1),\alpha_{j-1})$ is 0 when $\alpha_j = \alpha_{j-1}+1$. This encourages smooth transition among static scenes.

(3) We also assign positive cost to $C(\lambda(j),\alpha_j;\lambda(j-1),\alpha_{j-1})$ when $\alpha_j = \alpha_{j-1}$. This penalizes the decision of frame repeat. Therefore, a decision of frame repeat will be made only if it reduces the matching error significantly.

4. GENERALIZATION OF TEPORAL REGISTRATION

Our temporal video registration algorithm can be extended to include spatial and histogram registrations. This is done by incorporating them in the minimization of local prediction errors. We plan to present the spatial, temporal and histogram registration in another paper.

5. EXPERIMENTAL RESULTS

We have conducted the following experiment to test the performance and the robustness of our temporal registration algorithm. The videos that we use are standard definition movie clips with many scene changes. For these videos, we first distort them temporally. We randomly choose 20% of the frames to drop, 5% to repeat twice and 5% to repeat three times. 40% of frames are copied from the original to the processed sequences. For the last 30% of frames, we simulate frame integration using randomly generated λ's and linear combination of two consecutive frames to generate processed frames. We denote the temporally distorted sequences as "*TempDist*". Then, we further distorted the temporally distorted sequences as follows: (1) Gaussian low-pass filtering with standard deviation 1 and 3. The resulting videos are "*Gblur1*" and "*Gblur3*", respectively. *"Gblur3"* uses a 19x19 tap filters, and it introduces severe blur to a video. (2) Resizing: subsampling without anti-aliasing filtering, then interpolating back to the original resolution using bi-linear interpolation. We performed two resizing operations: 2x2 and 4x4 subsampling and interpolation. The resulting sequences are "*Resize2*" and "*Resize4*", respectively. Resizing not only introduces blurring, but also causes aliasing. "*Resize4*" also degrades the video quality significantly, similar to "*Gblur3*". (3) MPEG-4 compression at 2Mbps (Mega bits per second) and 1 Mbps. The results are "*MP4-2M*" and "*MP4-1M*". (4) Additive Gaussian noise with amplitude 5, denoted as "*Noise5*".

We registered all 8 types of processed videos to the original sequences. The results are shown in Table 1. Since the real values are known, we first computed the MAD (Mean Absolute Difference) of the estimate of λ_{ji} and listed them in the 3rd column in Table 1. The 4th column is the error rate (ER) of the estimate of $\alpha(j)$. If the estimated $\alpha(j)$ is not the same as the real $\alpha(j)$, we count it as one error. ER is the percentage of errors among all frames. There are only a few errors when the temporally distorted sequences are further distorted significantly, and all errors occurred in "Blur3" and "Noise5" have a difference of 1 between the real and the estimated $\alpha(j)$.

In addition, we show the effectiveness of our temporal registration using the RMSE (Root Mean Squared Error). The 5th column in Table 1 is the RMSE between the original and the processed sequences, the RMSE Before Registration. They are large because the temporal distortion causes misalignment among frames. The 6th column is the RMSE between the processed and the sequences that are only temporally distorted. This column shows the amount of additional distortion added to the temporal distortion. Since temporal registration can not compensate for these additional distortion. The RMSE in this column is the Lower Bound for the registration error. The last column is the RMSE between the processed videos and videos warped from the original videos using the results of the temporal registration, the RMSE After Registration. They are much smaller than the error between the original and the distorted sequences listed under "Before Reg" and very close to their "Lower Bounds" shown in the 6th column.

In conclusion, we propose a temporal registration algorithm for processed video sequences. Experimental results show that the proposed temporal registration algorithm is not only accurate, but also robust against various distortions caused by filtering, compression or additive noise.

Table 1. Experimental results discussed in section 6.

	Distorted Sequences	MAD λ_{ji}	ER %	RMSE		
				Before Reg	Lower Bound	After Reg
1	*TempDist*	0.00	0	49.9	0.00	0.70
2	*Blur1*	0.04	0	49.8	2.39	5.34
3	*Blur3*	0.11	1.7	49.5	6.05	13.23
4	*Resize2*	0.03	0	49.8	2.49	5.86
5	*Resize4*	0.06	0	49.7	5.02	11.11
6	*Noise5*	0.03	0.8	51.5	10.87	24.38
7	*MP4-2M*	0.03	0	49.8	2.48	5.50
8	*MP4-1M*	0.05	0	49.7	3.28	7.22

REFERENCES

[1] J. Lubin, J.A. Bloom and H. Cheng, "Robust second-generation watermarking for tracking in digital cinema," *Proc. of IS&T/SPIE Electronic Imaging*, Santa Clara, CA, Jan. 2003

[2] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, pp.163-178, Cambridge: MIT Press, 1993.

[3] Jiuhuai Lu, "Image analysis for video artifact estimation and measurement," *Proc. of SPIE Machine Vision Applications in Industrial Inspection*, v. 4301, pp. 166-174, San Jose, CA, Jan. 2001.

[4] Y. Caspi and M. Irani, "Alignment of non-overlapping sequences," *Proc. of IEEE Int'l Conf. on Computer Vision*, Vancouver, BC, Canada, July 2001.

[5] A.E. Bryson, *Dynamic Optimization*, Addison-Wesley, '98.

[6] C.S. Myers and S.E. Levinson, "Connected word recognition using a syntax-directed dynamic programming temporal alignment procedure," *ICASSP'81*, pp.956-9 vol.3. New York, NY, March 1981.