A CONFIDENCE MEASURE AND ITERATIVE RANK-BASED METHOD FOR TEMPORAL REGISTRATION

[1]*Meghna Singh,* [1]*Mrinal Mandal,* [2]*Anup Basu*

[1]Department of Electrical and Computer Engineering
[2]Department of Computing Science
University of Alberta, Edmonton, Canada.

ABSTRACT

In this paper we develop a confidence measure that can determine if a given set of samples is suitable for inclusion in the reconstruction of a higher resolution dataset. The confidence measure is formulated as a weighted combination of two well defined objective functions. We discuss the scope of the confidence measure and the two key factors that affect it: (i) non-uniformity of the samples and (ii) error in temporal registration. We also present a greedy iterative rank-based method that uses the confidence measure for reconstruction from multiple sample sets. The proposed method is evaluated with real video, audio and MRI data.

Index Terms— Temporal registration, Confidence measure, Video synchronization.

1. INTRODUCTION

In many applications, such as 3D rendering, super-resolution video, human activity recognition and video retrieval, acquiring multiple video streams can be beneficial. Unless these video streams are acquired via a hardware controlled timing mechanism, they are prone to temporal offsets caused by different acquisition start times or different frame rates of the cameras. Before fusing these video streams, the temporal offset between them needs to be computed. Various methods for computing temporal offsets have been proposed in the past and are referred to as temporal registration or video synchronization algorithms [1]-[5]. Some algorithms compute registration to sub-frame accuracy [4] while others perform a one-to-one frame correspondence [1]. However, most of these algorithms assume that the dynamics of the scene have been acquired at sampling rates higher than the Nyquist rate. This assumption is not true for applications in medical imaging (MRI) or videos of fast moving objects. When video acquisition is at low rates and less than the optimal number of samples are available, temporal registration algorithms report erroneous offsets. Also, the presence of noise in the videos results in error in feature computation which translates into error in the temporal registration. Since the original high-resolution event is unavailable for comparison, the degree of inaccuracy in the computed offset cannot be determined. In such conditions it would be useful to be able to compute (i) an estimate of how much confidence we have in the temporal registration and also (ii) an estimate of how much new information is added to the reconstruction process by the inclusion of a particular sample set.

Combining multiple video streams to generate a higher resolution video can also be formulated as a 2D case of recurrent non-uniform sample (RNUS) reconstruction, and therefore governed by the theorems associated with non-uniform sampling. *RNUS reconstruction* [7] refers to reconstruction of a signal from multiple sample sets which are offset from each other by an arbitrary time interval. Irrespective of whether we approach the fusion of video streams as a temporal registration problem or a RNUS problem, it is imperative to identify and discard video streams that will result in poor reconstruction. In this paper we present a measure to determine if a given set of samples is suitable for inclusion in the reconstruction of a higher resolution dataset. In [6], we presented a basic formulation of the confidence measure with preliminary experimental results. In this paper we extend our previous work to develop a generalized framework for computing the confidence measure based on two objective functions. We discuss the various factors influencing the confidence measure, from a temporal registration as well as RNUS point of view. We also develop a greedy rank-based method that relies on the proposed confidence measure to iteratively fuse multiple sample sets, while minimizing the reconstruction error. Experimental results with real and synthetic data are also presented.

2. PRELIMINARY DEFINITIONS

In this section we present some preliminary definitions that are required to formulate the confidence measure. Let S_i, where $i = 1..N$, denote N video sequences that are acquired (w.l.o.g.) at a constant frame rate and are offset from each other by a random time interval t_n such that the video sequences correspond to the 2D RNUS case discussed above. Each sequence S_i has M frames (I) such that $I_{i,k}$ denotes the k^{th} frame of the i^{th} video sequence (henceforth referred to as

sample set). Features (Ω) are extracted and tracked through all sequences to generate feature trajectories $\Omega_{i,k}$ (in the spirit of the discussion in [4]). In our past work [5], we built a continuous time event model from Ω, based on weighted linear least squares such that a continuous model of the feature space ($\Omega_{i,t}$) is estimated as follows:

$$\Omega_{i,t} = \Omega_{i,k}\beta_i + \epsilon_i, \tag{1}$$

where β_i is the regression parameter and ϵ_i is the model error term. We derive an estimate $\hat{\Omega}_{i,t} = \Omega_{i,k}\hat{\beta}_i$ by iteratively computing $\hat{\beta}_i$ such that a weighted residual error is minimized as follows:

$$minimize(\sum_{k=1}^{M} w_k \|\Omega_{i,k} - \hat{\Omega}_{i,t}\|^2 : t = k) \;\; w.r.t.\; \hat{\beta}_i. \tag{2}$$

Using event models results in a more accurate estimate of the subframe temporal offset compared to the commonly used linear interpolation approach [4]. Once the event models $\hat{\Omega}_{i,t}$ are available, the temporal offset (t_n) is computed using the following objective function:

$$t_n = [arg\min_{t_n} \sum_{i,j\in(1..N)} (\|\hat{\Omega}_{i,t} - \hat{\Omega}_{j,t+t_n}\|^2)]. \tag{3}$$

3. PROPOSED METHOD

While (3) is formulated as an objective function to compute the temporal offset between two sample sets, it can also provide information pertaining to how uniformly the samples are distributed and the minimum registration error achieved. We exploit this information to develop a generalized framework for computing the confidence measure. Generalizing the definition of local and global registration error from [6] we now define two objective functions Φ_g and Φ_l as follows (t_n is computed via (3)):

$$\Phi_g = \sum_{i,j,\in(1..N)} (\|\Omega_{i,kT} - \Omega_{j,kT+t_n}\|^2), \tag{4}$$

$$\Phi_l = \sum_{i,j,\in(1..N)} (\|\hat{\Omega}_{i,t} - \hat{\Omega}_{j,t+t_n}\|^2). \tag{5}$$

i.e., Φ_g is the minimum temporal registration error over the discrete samples acquired, while Φ_l is the minumum temporal registration error over the continuous event models of the samples. Note that in the case of video sequences, the sampling period T represents the inverse of the fixed frame rate of the cameras. In [6] we hypothesized that a large global registration error indicates a more uniform distribution of samples and a small local regsitration error represents a better temporal registration. Consequently, the following confidence measure is proposed:

$$\chi = w_g\Phi_g + w_l\Phi_l^{-1}, \tag{6}$$

where w_g and w_l are weights assigned to each of the objective functions. With no prior information on the sample sets, these weights can be set to 0.5. The proposed confidence measure (6) takes into account two factors that affect the reconstruction process: (i) the uniformity (or lack thereof) of sample data and (ii) the accuracy with which the datasets have been registered. In the following sections we justify why the above two factors play a pivotal role in defining the confidence measure and why Φ_g and Φ_l are suitable indicators of these two factors. We also present an iterative approach that uses χ in Eq. (6) to fuse multiple sample sets, providing a much better reconstruction performance, than an arbitrary fusion of sample sets.

3.1. Non-uniformity of Sample Sets

The central idea behind the non-uniform sampling theorem is that the non-uniform impulses of the sampling filter can be represented as a linear system of equations in terms of uniform sampling filter impulses. For example, suppose a signal $f(t)$ is sampled uniformly at t_0, t_1, t_2 and non-uniformly at t'_0, t'_1, t'_2, then using the Whittaker-Shannon Kotel'nikov (WSK) sampling theorem, the non-uniform samples can be represented in terms of the uniform samples as follows:

$$\begin{aligned} f(t'_i) = f(t_0)sinc(t'_i - t_0) + f(t_1)sinc(t'_i - t_1) \\ + f(t_2)sinc(t'_i - t_2) : i = (0, 1, 2). \end{aligned} \tag{7}$$

The above set of linear equations can be solved for sample values at the uniform sampling instances using methods such as LU decomposition, SVD and conjugate gradients. As long as the system of linear equations is not singular an interpolation formula for reconstruction from non-uniform samples can be formulated in the same manner as the WSK interpolation formula. The system of linear equations (7) can be solved if the equations are linearly independent, i.e.,

$$maximize[(t'_i - t_0) - (t'_j - t_0)] : i \neq j. \tag{8}$$

An interpretation of (8) is that for optimal interpolation or reconstruction, the recurrent sample sets should be as far away from each other as possible. If sample sets are maximally separated in time, then Φ_g (4) will be high. Thus, information about the non-uniformity of the sample sets can be derived from the objective function Φ_g.

3.2. Error in Temporal Registration

Computing temporal registration between sample sets is a non-trivial task. Since registration is computed as an optimization there is a possibility of error in the computation. Most algorithms in digital communications approach a similar problem of estimating sampling jitter by modeling the jitter as a Gaussian distribution. However, in our case we model the error in temporal registration $\delta(t)$ as a uniform distribution.

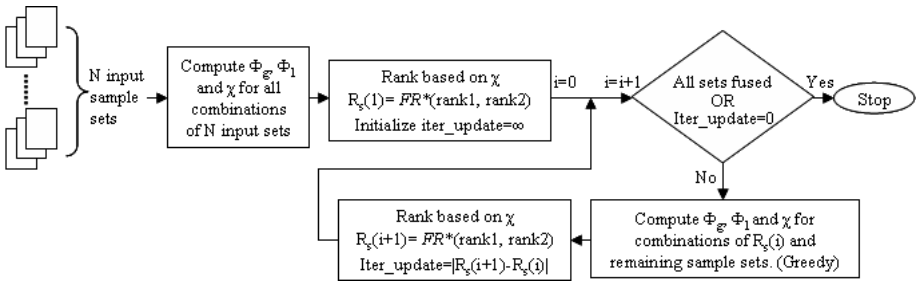

Fig. 1. Flowchart of iterative ranking method based on the proposed confidence measure. FR* indicates a RNUS fuse and reconstruction algorithm from [8].

This allows a more generic representation of the error and assumes no prior knowledge about the expected error other than a possible range. The computed temporal offset is therefore $t_n + \delta(t)$. Objective function Φ_l in Eq.(5) which sums the squared difference in the event models over the computed offset thus depends on $\delta(t)$. In other words, as the temporal registration estimate becomes more inaccurate, Φ_l increases. Experimentally we found the increase in Φ_l to be linearly related to the error in temporal registration. Φ_l also indicates that for a given distribution of error, there exists a threshold number of sample sets beyond which adding more sample sets is redundant and does not reduce reconstruction error. A mathematical formulation of such a threshold is beyond the scope of this paper, and will be dealt with in future work.

3.3. Iterative Rank-based Method

In reconstructing from multiple sample sets we need to order the sample sets such that the information added for reconstruction is maximized and the error in the reconstruction is minimized. This can be accomplished by ranking the multiple sample sets based on the computed confidence measures, as shown in Fig.1. We use ranking instead of directly using the numerical confidence measure scores as the scales of the confidence may change over each iteration, while ranking would be a more consistent relative measure of the confidence. Other implementations such as normalizing the confidence measure can also be used. We also assume that in each iteration the number of distinct ranks decreases by 1. In practice, however, confidence measure scores may result in ties. In such cases a weighted measure of the previous rank score can be added to the current rank to break the tie. The iterations are stopped when the difference between current reconstructed signal $R_s(i)$ and the signal from previous iteration $R_s(i-1)$ becomes small or when all the sample sets have been fused.

4. EVALUATION OF PROPOSED METHOD

We tested the rank-based reconstruction method on both synthetic as well as real data. Synthetic data was generated as a 1D high resolution random signal bandlimited to a user-controlled frequency. This high resolution data was then subsampled at random offsets to generate the multiple sample sets. Real data was collected via three sources: MRI videos of swallowing, real videos of a person swinging a ball and audio data from MATLAB demos.

Table 1. Results with real video sequences.

Scene	Sequence	χ	SSE
Scene 1	seq1-seq2	0	1700
	seq1-seq2	1	4.2
Scene 2	seq2-seq1	0	81200
	seq2-seq3	1	42000
Scene 3	seq3-seq1	0	20
	seq3-seq2	1	4
Scene 4 MRI	vid1-vid2	27.64	NA
	vid2-vid3	38.70	NA
	vid1-vid3	28.15	NA

For the experimental sample sets, we compute the proposed confidence measure and the rank the sample sets as per the algorithm in Section 3.3. The results of this approach with real video data, synthetic data and audio data are shown in Table 1 and Fig.2(a)-(b)respectively. Table 1(Scene 1-3) lists an absolute value of 1 or 0 for the confidence measure χ and also the corresponding sum of squared error (SSE). It can be seen that $\chi = 0$ (low confidence) indeed corresponds to a high SSE. Results corresponding to Scene-4 MRI are discussed later in this section. Figure 2(a) plots the normalized reconstruction error versus the number of sample sets added during reconstruction, for synthetic test data. It can be seen that if the sample sets are arbitrarily chosen and fused, the reconstruction error is much higher than if the sample sets are ranked and chosen based on the proposed confidence measure. In some cases, the proposed method resulted in lower reconstruction error with a few ranked sample sets than even all sample sets combined, as illustrated by point-1 and point-2 in Fig.2(a). For audio data, we use a section of an audio signal available in the MATLAB demo data as 'toilet.wav.' We assume that low resolution audio data is available and multi-

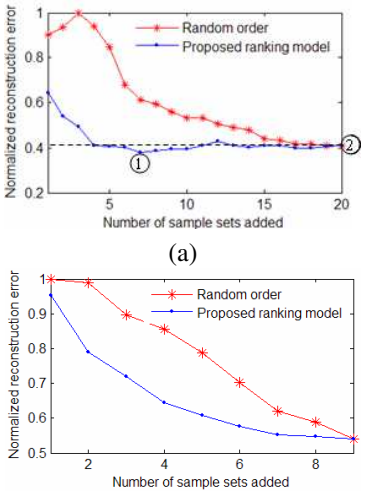

Fig. 2. Results of the iterative ranking method for (a) synthetic and (b) audio data.

ple such low resolution files are generated by subsampling the original wavefile. The sampling rate is below the Nyquist rate. Sample sets are temporally aligned and reconstructed based on the proposed confidence measure and ranking system. Results are shown in Fig.2(b). It can be seen that the proposed confidence measure and the ranking system successfully order the audio sample sets such that lesser number of sample sets are needed to reconstruct the same signal, compared to an arbitrary ranking of the audio sample sets.

We also acquired three MRI videos of a person swallowing a fixed amount of water. With MRI data, ground truth registration information is not available (SSE w.r.t. to the ground truth cannot be computed), hence we visually determine the registration. The original as well as temporally registered and reconstructed MRI videos can be viewed at: *www.ece.ualberta.ca/ ˜ meghna/ICASSP08.html*. Visually it is seen that fusing vid2-vid3 results in the best registration and reconstruction, see Fig.3. Confidence measures computed between vid1-vid2, vid2-vid3 and vid1-vid3 are listed in Table 1 Scene-4. It can be seen from the results that the confidence measure for vid2-vid3 combination is indeed the highest. Thus, our iterative ranking system, which uses the generalized confidence measure, performs much better than an arbitrary ordering of the sample sets during reconstruction, in both synthetic and real test cases.

5. CONCLUSION AND FUTURE WORK

In this paper, we developed a confidence measure that allows us to choose between recurrent non-uniform samples such that the overall signal reconstruction error is minimized. The confidence measure is based on two objective functions - (i) Φ_g which indicates the non-uniformity in sampling, and, (ii)

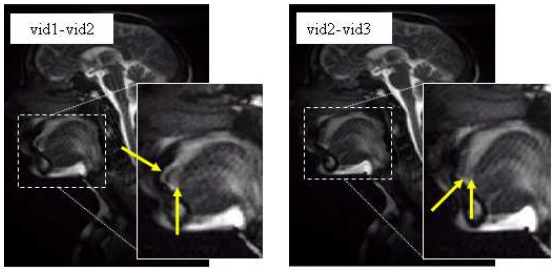

Fig. 3. Results of registered and fused MRI. (a) vid1-vid2, $\chi = 27.64$, zoomed position of the tongue shows incorrect registration, (b) vid2-vid3, $\chi = 38.7$, zoomed position of tongue shows correct registration.

Φ_l which indicates the reliability of the temporal registration. We also develop a ranking system that iteratively updates the rank assigned to sample sets and fuses them to optimize reconstruction. Such a ranking system based on the confidence measure is shown to outperform an arbitrary ordering of the sample sets, which would otherwise be used when no prior information about the sample set order is known. In the future, we would like to evaluate the assignment of the weights to the objective functions and develop a strategy to tune these weights for event specific sample sets.

6. REFERENCES

[1] C. Rao, A. Gritai, M. Shah and T. Syeda-Mahmood, *View-invariant alignment and matching of video sequences*, Proc. Intl. Conf on Computer Vision, vol.2, Oct 2003, pp 939-945.

[2] P. Tresadern and I. Reid, *Synchronizing image sequences of non-rigid objects*,Proc. British Machine Vision Conference, Norwich, vol.2, Sept 2003, pp 629-638.

[3] R.L. Carceroni, F.L.C. Padua, G.A.M.R. Santos and K.N. Kutulakos, *Linear sequence-to-sequence alignment*, Proc. Intl. Conf. on Computer Vision and Pattern Recognition, July 2004, pp I-746 - I-753.

[4] E. Shechtman, Y. Caspi and M. Irani, *Space-time super-resolution*, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 27, No.4, April 2005, pp 531-545.

[5] M. Singh, A. Basu and M. Mandal, *Event dynamics based temporal registration*, IEEE Trans. on Multimedia, vol.9, no.5, Aug. 2007, pp 1004-1015.

[6] —-, *Confidence measure for temporal registration of recurrent non-uniform samples*, Proc. Intl. Conf. on Pattern Recognition and Machine Intelligence, Kolkatta, Dec 2007.

[7] H. Freeman, *Discrete-Time Systems*, John Wiley and Sons Inc., 1965.

[8] F. Marvasti, (ed.), *Nonuniform Sampling Theory and Practice*, New York: Kluwer Academic, 2001.