Synchronization Using Shapes

Richard Chang Sio-Hoi Ieng Ryad Benosman
Institut des Systemes Intelligents et de Robotique
Universite Pierre et Marie Curie, Paris6
4 Place Jussieu 75252 Paris,France
richard.chang@isir.fr

Abstract

The synchronicity is a strong restriction that in some cases of wide applications can be difficult to obtain. This paper studies the methodology of using a non synchronized camera network. We consider the cases where the frequency of acquisition of each element of the network can be different, including desynchronization due to delays of transmission inside the network. The following work introduces a new approach to retrieve the temporal synchronization from the multiple unsynchronized frames of a scene. The mathematical characterization of the 3D structure of scenes, is used as a tool to estimate synchronization value, combined with a statistical stratum. This paper presents experimental results on real data for each step of synchronization retrieval.

1 Introduction

The synchronization operation is a task that complexifies many vision operations as the number of cameras becomes higher : cameras calibration, 3D reconstruction, frames synchronization, etc... Baker and Aloimonos [2], Han and Kanade [8] introduced pioneering approaches of calibration and 3D reconstruction from multiple views. The reader may refer to [13, 14, 5] for other interesting work on camera networks. The aim is to retrieve synchronization in order to compute correctly 3D structures from a set of cameras. A solution is to set hardware synchronization as in [10]. But this kind of method cannot be appliable because of spatial constraints. In these cases, a software based synchronization can be a way to solve this problem. Most of the former works assume cases of desynchronization with highly constraints hypotheses which exclude heavy and non linear delays [19]. In [15, 16], a set of five moving points is tracked and matched throughout sequences for synchronization. Constraints can also be set on the scene or on the geometry of the cameras studying feature points [13] or trajectories [4] of the objects. Ushizaki et al. [17] show the limitations of these approaches and present a method based on co-occurrences of appearance changes in video sequences. This method uses appearance changes as temporal features but it may fail when appearance changes, due to temporal shift and the cameras have to be stationary. In this paper, we introduce a new synchronization technique from 3D structures. From all available frames which can be synchronized or not, 3D structures are computed regardless they are correct or not. We will show that correct ones are only generated from synchronized frames. If we have a prior knowledge of the exact models of the observed objects, synchronization can be recovered by determining frames that lead

BMVC 2008 doi:10.5244/C.22.75

to shapes complying with the models. However most of the time, this knowledge is not available. We then introduce a statistical approach which assumes that correct shapes reconstructions (computed from synchronized frames) occur more frequently than distorted ones (computed from non synchronized frames). A distribution model of the 3D reconstructions can be established where wrong shapes are marginal cases of the correct ones. We will also explain the method used to compute 3D shapes from available frames and the way we characterize them such that discrimination between correct and wrong reconstructions is possible. This paper is organized as follows. Section two describes the theoretical fundaments of our method of synchronization. In section three, several shape characterizations used to classify reconstructed shapes are examined. In section four a propagation method is introduced and finally the section five presents experimental results of the synchronization of a camera network.

2 Problem formalization

2.1 Shape criterion for synchronization.

It is reasonable to assume that correct reconstructions are possible if frames are synchronized and that unsynchronized frames lead likely to distorted results. We will prove in this section that this assumption is mathematically true : "correct reconstructions" are equivalent to "synchronized frames" if observed objects are rigid bodies. This can be done by examining simple planar motions.
Let $\mathbf{P}_1$,$\mathbf{P}_2$,$\mathbf{P}_3$ and $\mathbf{P}_4$ be four collinear points viewed by C_R and C_L of centers $\mathbf{O}_L$ and $\mathbf{O}_R$ (see figure 1). Since the $\mathbf{P}_i$ are collinear, we have the following relations :

$$\mathbf{P_1P_2} = K\mathbf{P_1P_4} \quad \text{and} \quad \mathbf{P_3P_2} = M\mathbf{P_3P_4} \tag{1}$$

where K and M are constant scalars and we define $\mathscr{L} = ||\mathbf{P_1P_4}||$. We assume that the structure of the points is known only in this section for demonstration purpose. In the rest of the paper, the shape of the objects is unknown and to be determined.

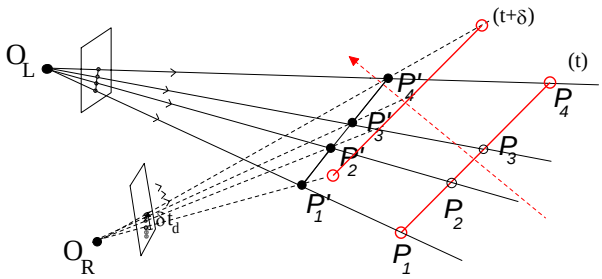


Figure 1: If the images from the cameras C_R and C_L are synchronized, the points $\mathbf{P}_1$, $\mathbf{P}_2$, $\mathbf{P}_3$ and $\mathbf{P}_4$ can be correctly triangulated from images. If not, the triangulation produces shifted $\mathbf{P}'_1\mathbf{P}'_2\mathbf{P}'_3\mathbf{P}'_4$ at different positions according to the rigid body hypothesis.

When the cameras C_R and C_L are synchronized, we have a correct 3D reconstruction and the relations in eq.(1) are always satisfied whether the structure is moving or not.

If the cameras are not synchronized, the rays will produce a new point set $\{\mathbf{P}'_i\}$ which is different to the set $\{\mathbf{P}_i\}$ (see figure 1). Since we only assume non deformable body, if the collinearity is not preserved by the $\mathbf{P}'_i$ then the reconstructions are obviously wrong, thus we are only considering cases for which the $\mathbf{P}'_i$ are collinear. In such condition we can similarly establish relations as eq. (1) with $\mathscr{L}'$, K' and M' for these points. The $\mathbf{P}'_i$ are incorrectly reconstructed points if some trivial metric properties satisfied by the $\mathbf{P}_i$ are no longer true. Then, we can apply the cross ratio between the different lines:

$$\frac{\| \mathbf{P_1P_2} \|}{\| \mathbf{P_1P_4} \|} / \frac{\| \mathbf{P_3P_2} \|}{\| \mathbf{P_3P_4} \|} = \frac{\| \mathbf{P'_1P'_2} \|}{\| \mathbf{P'_1P'_4} \|} / \frac{\| \mathbf{P'_3P'_2} \|}{\| \mathbf{P'_3P'_4} \|} \tag{2}$$

If $K = K'$, then $M = M'$ hence :

$$\frac{\| \mathbf{P_3P_2} \|}{\| \mathbf{P_3P_4} \|} = \frac{\| \mathbf{P'_3P'_2} \|}{\| \mathbf{P'_3P'_4} \|} \tag{3}$$

This equality is the Thales' theorem, satisfied by $\mathbf{P}_i$ and $\mathbf{P}'_i$ only if the lines $(\mathbf{P}_1\mathbf{P}_4)$ and $(\mathbf{P}'_1\mathbf{P}'_4)$ are parallel. If $K = K'$, then there is only one reconstruction that also satisfies $\mathscr{L} = \mathscr{L}'$. This solution corresponds to the case where $\mathbf{P}'_i = \mathbf{P}_i$ (the case where the points are behind the centre of camera is rejected). This proves that for non synchronized cameras, the exact reconstructions of simple rigid structures are not possible, thus we can expect better result for complex ones.

2.2 Using recurrence for correct shapes extraction

The correctness of the reconstruction provides a good criterion to recover synchronization between cameras. Recurrent shapes can be reliably used to sort out correct structures from bad ones if there is no ambiguity. We examine here if desynchronizations can produce enough recurrent wrong shapes of length $\mathscr{L}'$ that can compete with those corresponding to $\mathscr{L}$.

We assume δ as the temporal shift between C_L and C_R and $\mathbf{O}_L$ is chosen as the origin of the world coordinate frame. The $\mathbf{P}_i$ define an object moving through the scene (figure 2). C_L sees $\mathbf{P}_1$ at t (i.e.$\mathbf{P}_1(t)$) and because of δ, C_R will see the same point at $t+\delta$ (i.e. $\mathbf{P}_1(t+\delta)$). The reconstruction $\mathbf{P}'_1$ of $\mathbf{P}_1$ from these frames will be the intersection of $(\mathbf{O}_L\mathbf{P}_1(t))$ and $(\mathbf{O}_R\mathbf{P}_1(t+\delta))$, satisfying :

$$\mathbf{P}'_1(t) = \alpha_1\mathbf{P}_1(t) = \mathbf{O}_R + \alpha'_1(\mathbf{P}_1(t+\delta) - \mathbf{O}_R) \tag{4}$$

where $(\alpha_1, \alpha'_1) \in \mathbb{R}^2$. This equality is a set of three equations from which the scales factors can be expressed from the other parameters. By combining them, we can express α_1 with the known parameters :

$$\alpha_1 = \frac{det(\mathbf{P}_1(t+\delta) - \mathbf{O}_R, \mathbf{O}_R)}{det(\mathbf{P}_1(t+\delta, \mathbf{P}_1(t))} \tag{5}$$

Similar equations can be established for $\mathbf{P}_4$, hence L', the norm of the $\mathbf{P}'_1\mathbf{P}'_4$ can be expressed as a function of $\mathbf{P}_1$ and $\mathbf{P}_4$:

$$\mathscr{L}' = \| \mathbf{P}'_4 - \mathbf{P}'_1 \| = \| \alpha_4\mathbf{P}_4 - \alpha_1\mathbf{P}_1 \| \tag{6}$$

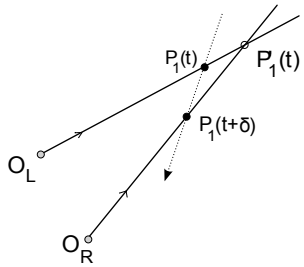


Figure 2: Due to the delay, the point $\mathbf{P}'_1$ is constructed from points $\mathbf{P}_1(t)$ and $\mathbf{P}_1(t+\delta)$ seen by C_L and C_R.

We assume now that the value of L' is set and given a neighborhood $\mathscr{D}$ of the cameras, we look inside it for all $\mathbf{P}_1$ and $\mathbf{P}_4$ that produce $\mathbf{P}'_1\mathbf{P}'_4$ of length $\mathscr{L}'$. This is done by minimizing the cost function with respect to $\mathbf{X} = \begin{bmatrix}\mathbf{P}_1 & \mathbf{P}_4\end{bmatrix}^t$:

$$E(\mathbf{X}) = \left(||\alpha_4\mathbf{P}_4 - \alpha_1\mathbf{P}_1|| - \mathscr{L}'\right)^2 \tag{7}$$

Equation 7 is solved for several values of $\mathscr{L}'$ and for several initial conditions. However the recovered lengths' dispersion is too high to satisfy any length preservation (60 % around the mean value).

3 Shape characterization

3D reconstructions provide information for synchronizing the cameras. Shapes being at the core of the method, it is compulsory to set characterization of reconstructed structures in order to compare and classify them. 3D shapes characterization has been intensively studied in indexation techniques [11, 7]. Most of these approaches establish a mapping between a 3D object and some vector space of dimension n so that each object can be summarized by a vector of n components defined as a signature. In our case, we are using two kind of mappings : one is based on decomposition of the object into spherical harmonics and the other one is based on the distribution of the distance between two randomly selected points on the object surface [11]. We set the following notations:

- S is an object moving through the scene viewed by m cameras.
- f defines the size of the search interval F.
- S_n^j is the reconstruction computed from the j^{th} image combination included in F. This interval F is centered on the n^{th} image of an arbitrarily chosen camera C_1.

The S_n^j are geometric reconstructions obtained with voxel coloring method [12]. This method has many advantages as it gives a dense reconstruction of objects, and is easy to compute. We then compute these 3D reconstructions for the j combinations of images defined for every image n according to the search interval F. For each S_n^j, both mentioned classification techniques are applied.

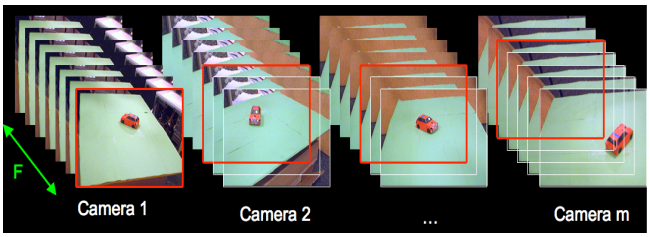


Figure 3: For each frame of C_1, an interval of length f is set. This defines locally a set of images acquired by the m cameras, reconstructions are then performed by combining frames from each camera.

Critical Motion One of our main hypothesis for synchronization recovery is the use of motions. Static objects will not allow discrimination between synchronized and non synchronized frames since correct reconstructions are possible whatever the timeshifts are. The same results will occur for stationary motions. That is the motions combined to the delays between frames produce globally invariant projections in the images planes.

The method can also deal with heavy time delays. In the general case, the temporal window is initialized large enough but is readjusted according to the retrieved synchronization. The number of reconstructions is high at the beginning but decreases over time.

3.1 Characterization with spherical harmonics

As the equivalent of the Fourier series for 3D functions, the spherical harmonics decomposition is suitable to express 3D shapes with a set of orthogonal functions. Funkhouser et al.[6] and Vranic[18] used it to characterize 3D shapes for indexation and model retrieval. Given a 3D object $\mathscr{S}$ defined as a surface points set $\{\mathbf{X}\}$, we sample it according to a $2R \times 2R \times 2R$ voxel grid. The object is normalized and scaled around its gravity center according to :

$$\mathscr{S}' = \left\{ \mathbf{X}' \in \mathbb{R}^3 | \mathbf{X}' = \frac{\mathbf{X} - (R,R,R)^t}{R/2} \right\} \tag{8}$$

The spherical decomposition is then applied to $\mathscr{S}'$ according the following assumptions : $\mathscr{S}'$ is sampled regularly by intersecting it with n concentric spheres Ω_k of radii r_k (and $r_n = R$), centered on $(R,R,R)^t$. For each Ω_k, we define the function $f_k(\theta,\phi)$ as $\mathscr{S}' \bigcap \Omega_k$. If the Fourier transform is applied on f_k over the sphere Ω_k, we have :

$$f_k(\theta,\phi) = \sum_m f_k^m(\theta,\phi) = \sum_m \sum_{n=-m}^{m} a_{mn} \sqrt{\frac{(2m+1)}{4\pi} \frac{(m-|n|)!}{(m+|n|)!}} \times P_{mn}(cos\theta) e^{in\phi} \tag{9}$$

where the P_{mn} are the Legendre functions (1st kind) of degree m and order n, and the a_{mn} are the Fourier coefficients. f_k^m can be interpreted as the projection of f_k on the m^{th} representation of rotation group in harmonic spherical space. We then form the rotation invariant signature s_k of f_k with the L first spherical harmonics and the feature vector $\mathbf{V}_{SpH}$ is computed for each 3D reconstruction S_n^j as its signature vector:

$$\mathbf{V}_{SpH} = (s_0, s_1, \ldots, s_{n-1}) \text{ where } s_k = (|f_{k,0}|, |f_{k,1}|, |f_{k,2}|, \ldots, |f_{k,L-1}|) \tag{10}$$

3.2 Characterization with distance distribution

This method is based on a statistical consideration of the geometric properties of a 3D shape (see [1]). It establishes the distribution of distances between 2 randomly selected surface points of the object. We form its distances histogram which is normalized into a feature vector $\mathbf{V}_d$. Due to the low complexity of this technique, the required computational load is limited. These characterization vectors allow the computation of the shapes distribution with respect to a reference geometric structure (normalized sphere for instance). More precisely, the distribution of the distances $d(\mathbf{V}, \mathbf{V}_{\text{ref}})$ of each vector to the reference structure's one is computed ($d()$ is any suitable distance e.g. Euclidean, Minkowski, ...). From this distribution, one can identify the most recurrent signature, hence the correct shape.

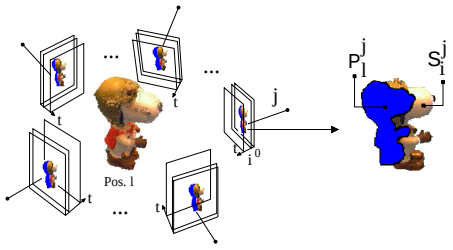


Figure 4: Synchronization propagation. The 3D model reinserted into the scene is projected as $S_i^{'}$ in the i^{th} the camera. If the coherence is maximum, the camera is synchronized.

4 Optimization : extension to large networks

Computational time is the major limitation of the synchronization method : we have to perform 3D reconstructions for each combination of images in the temporal window. As the number of cameras increases, the computational load becomes quickly unacceptable. In order to reduce it, we propose to split the synchronization into two parts as in [3]: (1) We first synchronize a small subset of cameras of the network according to our technique. Hence this subset is able to provide the correct reconstruction of any scene structure. (2) Assuming now that all cameras are calibrated and observing an object which correct 3D model is provided by the synchronized subset, we can propagate the synchronization by comparing the projections of both the model and the real object. A camera is synchronized if the projections are equal in the sense of they exactly overlap each other.

Let S_i and $S_i^{'}$ be respectively the silhouette of the object and the one of its 3D model in the i^{th} camera (see figure 4). We compute for each camera the coherence C as defined in [9] :

$$C(S_i, S_i^{'}) = \frac{\int (S_i \cap S_i^{'})}{\int S_i} \tag{11}$$

A desynchronization will produce shift between both silhouettes, the quantity C is a decreasing function as the desynchronization increases. The i^{th} camera can then be synchronized if C is maximized.

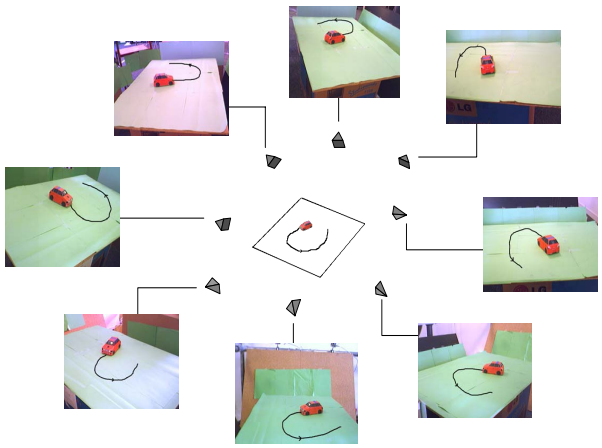


Figure 5: (a) A set of 8 cameras placed around the scene, watching an object moving inside. The reconstructions as previously stated and their characterization vectors are computed.

5 Experimental results.

5.1 Synchronizing an eight cameras network

The presented method is applied to synchronize a set of eight cameras placed all around a scene inside which a rigid body is moving (see figure 5). The desynchronizations between the video signals result from cumulation of hardware latencies : non equal cameras starting time, signals integration which is not instantaneous, data transfert time, etc...

Given the videosequences acquired by the cameras, 3D shapes are reconstructed by using frames defined within the search interval F. The characterization vector is established for each reconstruction and used to compute the shape distribution with respect to a unit sphere. Figure 6 shows the different normalized measures of distances of each shape to the unit sphere. Since the correct shapes are also the most recurrent, the characterization vectors should be stable hence the standard deviation of their distances is minimal.

If correct shape is recovered, its correct trajectory can also be extracted. The estimated trajectory is compared to ground truth data given by a camera observing the scene from above and locating the observed object using a visual tag. The results are shown by figure 7 and underline the accuracy of the approach.

5.2 Synchronization propagation.

The synchronization is extended to a twenty-four camera network according to the propagation techniques described in section 4 from a subset of six synchronized cameras. The object is a person moving inside the scene whose upper body part is assumed to be partially not deformable. The 3D model is computed from a subset of cameras synchronized with the method and is used to compute the coherence of the silhouettes for each unsynchronized camera. A led panel provides time ground truth by identifying frame in each stream that shows the start time of the panel. The figure 8(a) shows the 3D model of the

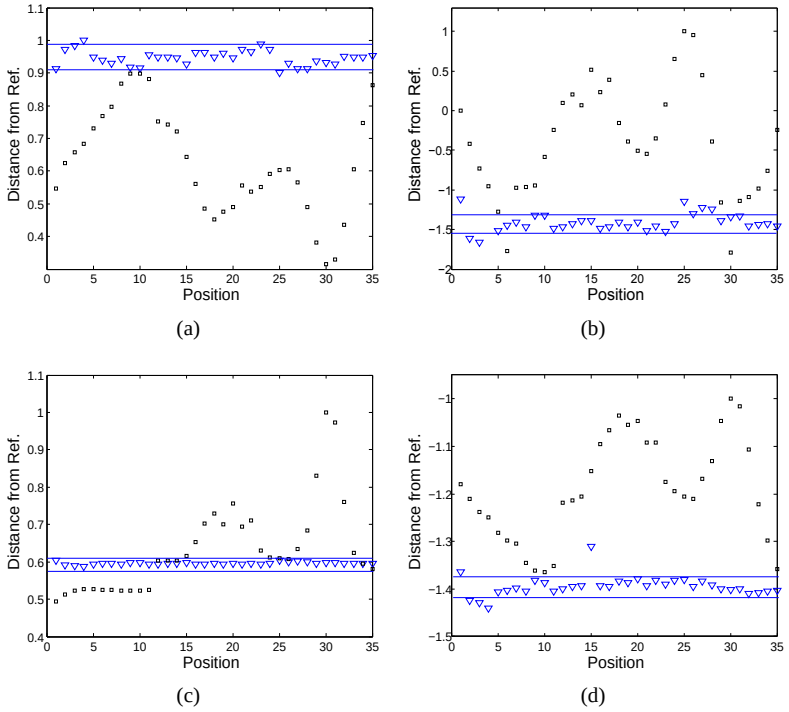


Figure 6: Normalized distances measurements between each characterization vector and the unit sphere : (a) Euclidean, (b) Minkowski, (c) Intersection and (d) Bhattacharrya. The correct shapes can be detected as the populations, represented here by the triangles, with minimum standard deviation.

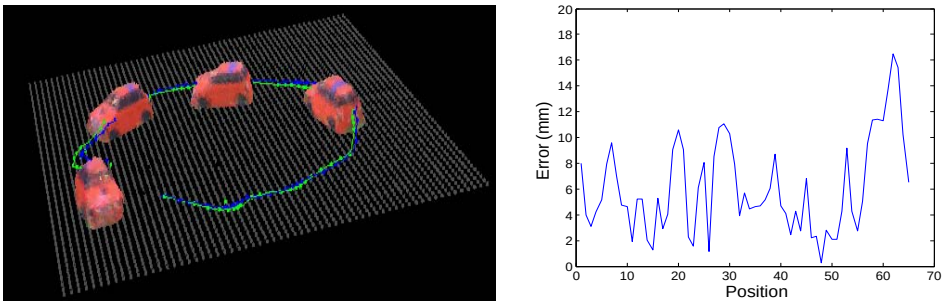
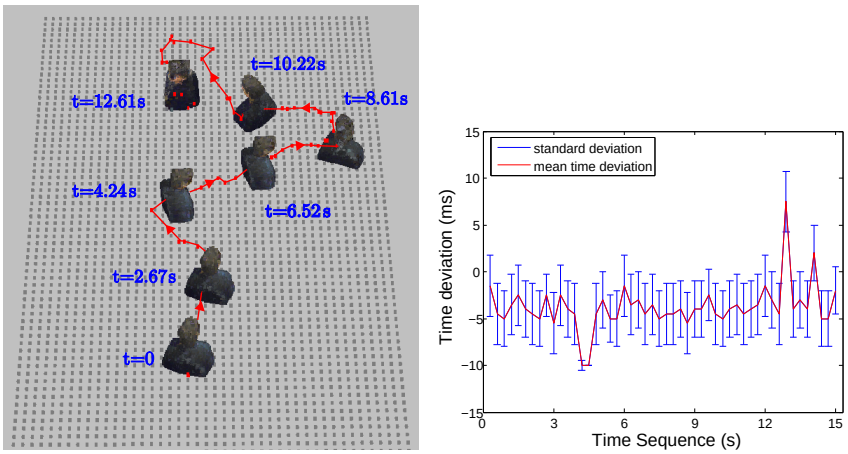


Figure 7: (a) Trajectories computed from the correct shapes with the synchronization method and the ground truth one. (b) For each position, both curves are compared to each other by measuring a ”point-to-point” error. The mean error (5mm) is reasonably small enough compared to the magnitude of the trajectory ($\sim 35cm$, the recovered trajectory error is less than 2%).



(a) (b)

Figure 8: (a) Reconstructed upper body along its trajectory. A precise datation is available with to the ground truth given by the led panel. (b) Mean delays measured from the synchronized frames using the led panel reading. Since the mean deviation is 5ms for a standard deviation of 3ms, the accuracy of the datations is up to 10^{-2}s.

entire trajectory of the man inside the network after the computational synchronization. A mean datation can then be assigned to several positions of the man, computed from the frames used to build the correct 3D shapes. For a total of 52 correct reconstructions through the man's motion, the mean relative delay and the standard deviation of all cameras (with respect to the panel reading) are represented in figure 8(b). As one can see, after the propagation process, the relative desynchronization is approximately equal to 5ms for a mean standard deviation of 3ms. Finally we also have to mention that if cameras framerates that are set to n fps, then for each individual sensor, the desynchronization (with respect to the global clock) accuracy cannot be higher than $\frac{1}{n}$ second, however the relative mean desynchronization of an entire network accuracy can be below this limit due to the contribution of each stream.

6 Conclusion

We proposed in this paper a new method to synchronize a set of cameras. We proved the possibility to recover the time shifts between the cameras from scene structures without need of any external hardware. The constraints set on the scene are limited to the hypothesis of mobile rigid bodies. If our method can benefit from a prior knowledge of the geometric models of the bodies to recover the synchronization, it can also provide solution in more general cases where such an information is not available. We also showed the equivalence between synchronization and correct structures reconstructions. In order to reduce computational loads as the number of camera increases we introduced a propagation process to synchronize a large network with a small subset of already synchronized cameras. Unnecessary reconstructions can then be avoided.

References

[1] M. Ankerst, G. Kastenmller, H.-P. Kriegel, and T. Seidl. 3d shape histograms for similarity search and classification spatial databases. In *SSD*, 1999.

[2] P. Baker and Y. Aloimonos. Complete calibration of a multi-camera network. In *Omnivis*, 2000.

[3] Joao Barreto and Kostas Daniilidis. Wide area multiple camera calibration and estimation of radial distorsion. In *Omnivis*, 2004.

[4] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence to sequence matching. In *IJCV*, 2007.

[5] J. Domke and Y. Aloimonos. Multiple view image reconstruction: A harmonic approach. In *CVPR*, 2007.

[6] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3d models. In *ACM Trans. Graph.*, 2003.

[7] Hironobu Gotoda. 3d shape comparison using multiview images. *National Institute of Informatics Journal*, 2003.

[8] M. Han and T. Kanade. Creating 3d models with uncalibrated cameras. In *IEEE Workshop ACV*, 2000.

[9] C. Hernandez, F. Schmitt, and R. Cipolla. Silhouette coherence for camera calibration under circular motion. In *PAMI*, 2007.

[10] G. Litos, X. Zabulis, and G. Triantafyllidis. Synchronous image acquisition based on network synchronization. In *CVPR*, 2006.

[11] Osada, T. Funkhouser, B. chazelle, and D. Dobkin. Shape distributions. In *ACM Trans. Graph.*, 2002.

[12] S. Seitz and C. Dyer. Photorealistic scene reconstruction by voxel coloring. In *IJCV*, 1999.

[13] S. Sinha and M. Pollefeys. Synchronization and calibration of camera networks from silhouettes. In *ICPR*, 2004.

[14] T. Svoboda, D. Matinec, and T. Pajdla. A convenient multi-camera self-calibration for virtual environments. In *PRESENCE*, 2005.

[15] P. Tresarden and I. Reid. Synchronizing image sequences of non-rigid objects. In *BMVC*, volume 2, 2003.

[16] T. Tuytelaars and L.J. Van Gool. Synchronizing video sequences. In *CVPR*, 2004.

[17] Manabu Ushizaki, Takayuki Okatani, and Kochiro Deguchi. Video synchronization based on co-occurrence of appearance changes in video sequences. In *ICPR '06*.

[18] D. V. Vranic. *3D Model Retrieval*. PhD thesis, University of Leipzig, 2004.

[19] A. Whitehead, R. Laganiere, and P. Bose. Temporal synchronization of video sequences in theory and in practice. In *IEEE Workshop on MVC*, 2005.