A Robust Temporal Alignment Technique for Infrared and Visible-Light Video Sequences

Andreas Ellmauthaler, Eduardo A. B. da Silva, Carla L. Pagliari and Jonathan N. Gois

***Abstract*— Traditional temporal video alignment techniques depend on the existence of a sufficient number of mutual scene characteristics to perform temporal synchronization. However, for video sequences originating from different spectral modalities such information may fail to exist, subsequently leading to unreliable alignment results. In this work, a novel, practical approach for the temporal alignment of infrared (IR) and visible-light video sequences is introduced. Utilizing a planar alignment device consisting of an array of miniature light bulbs, we propose to increase the number of mutual feature points within the frames of a temporally misaligned IR/visible-light video sequence pair. Thereby, we provide an abundant number of mutual scene characteristics which can subsequently be used to determine the time-shift between the two sequences. Tests with five different IR/visible-light sequence pairs suggest that our technique is able to estimate the temporal offset with a very high confidence level. Furthermore, due to the simplicity of the construction of the alignment board, the proposed framework can be seamlessly integrated within a stereo camera calibration scheme.**

***Keywords*— temporal video alignment, video synchronization, infrared imaging, feature point extraction**

***Resumo*— Técnicas tradicionais de alinhamento temporal de vídeos dependem da existência de um número suficiente de características mútuas nas cenas para realizar a sincronização. Entretanto, para sequências de vídeos de diferentes origens espectrais, este tipo de informação pode não existir, consequentemente levando a um alinhamento impreciso. Neste trabalho uma nova abordagem prática para alinhamento temporal de sequências de vídeos infravermelho (IR) e de luz visível é introduzida. Utilizando um equipamento planar de alinhamento composto por uma grade de lâmpadas em miniatura, podemos aumentar o número de características mútuas entre os quadros dos pares de vídeos IR/luz visível desalinhados temporalmente. Assim, nós obtemos uma quantidade suficiente de características para a sincronização, que pode ser posteriormente utilizada para determinar o deslocamento temporal entre as duas sequências. Testes com cinco diferentes sequências de pares de vídeos IR/luz visível sugerem que nossa técnica é capaz de estimar o deslocamento temporal com um alto grau de confiabilidade. Além disso, devido à simplicidade de construção do padrão de alinhamento, a estrutura proposta pode ser usada para a calibração de câmeras estéreo.**

***Palavras-Chave*— alinhamento temporal de vídeos, sincronização de vídeos, imagens infravermelhas, extração de pontos características**

Andreas Ellmauthaler, Eduardo A. B. da Silva and Jonathan N. Gois are with the Universidade Federal do Rio de Janeiro, PEE/COPPE/DEL, Rio de Janeiro, Brazil, E-mail: {andreas.ellmauthaler, eduardo, jonathan.gois}@smt.ufrj.br; Carla L. Pagliari is with the Instituto Militar de Engenharia, Department of Electrical Engineering, Rio de Janeiro, Brazil, E-mail: carla@ime.eb.br

I. Introduction

Temporal alignment of multiple video sequences is the preliminary step in many computer vision applications such as stereo camera calibration, 3D scene reconstruction, robot navigation or image fusion. It is defined as finding the offset (or time-shift) between a reference video and several other video sequences such that the final, temporally aligned video sequences exhibit the highest similarity possible. Typically, such time-shifts occur when the employed cameras are not activated simultaneously and/or have different frame rates. Note that in some applications it is feasible to synchronize cameras using hardware. Nevertheless, this is not practical in situations where the cameras are physically separated or mobile. In general, the individual sequences are recorded from distinct, but stationary, viewpoints showing the same scene. However, they may exhibit different zoom factors and image resolutions which may turn the exact temporal alignment challenging.

Temporal alignment techniques can roughly be classified into two groups: feature-based approaches and so-called direct methods. Feature-based methods [1], [2], [3], [4], [5], [6] rely on the detection of a sufficient number of feature points along the source images belonging to the same video sequence. Based on these feature points, methods such as RANSAC [7] are employed to establish correspondences between the sequences and to subsequently calculate the temporal offset. In contrast, so-called direct methods [1], [8], [9] perform temporal alignment by exploiting common scene characteristics within the input videos (e.g. common illumination changes, appearance/disappearance of an object present in all video sequences).

In general, both feature-based and direct methods are well-suited for the temporal alignment of video sequences originating from cameras operating in the same spectral band. However, they tend to face problems for sequences obtained by sensors of different modalities (such as infrared (IR) and visible-light sensors). For feature-based methods this is mainly due to the lack of mutual interest points between the IR and visible-light image sequences. An analog problem appears when employing direct methods. For instance, illumination changes tend to appear solely in the visible-light video sequence and not in the IR sequence.

In this paper a novel approach for the temporal alignment of IR and visible-light video sequences is introduced. Motivated by the frequently occurring lack of mutual scene characteristics between IR and visible-light video sequences,

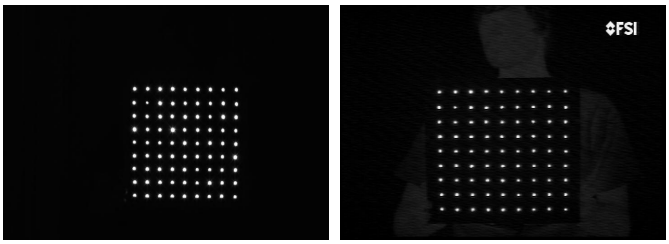

Fig. 1. Employed alignment board consisting of 81 light bulbs, arranged in a 9×9 matrix, in the visible-light (left) and IR spectrum (right). The depicted images were taken from an IR/visible-light image sequence pair after temporal alignment.

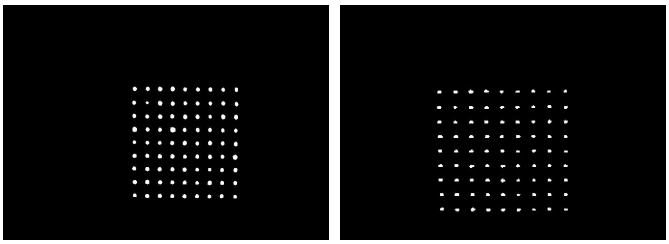

Fig. 2. Result of the adaptive thresholding operation when applied to Fig. 1. (Left) Binarized visible-light image. (Right) Binarized IR image.

we propose to employ a planar board (henceforth referred to as alignment board) equipped with miniature light-bulbs to increase the number of mutual feature points (light-bulb locations) between both recorded video sequences. The extracted feature point locations along both sequences are subsequently used to estimate the time-shift between the IR and visible-light sequences. This setup is of special interest since the a-priori knowledge of the physical properties of the alignment board allows for a precise localization of the light bulb positions, yielding very robust temporal alignment results. Furthermore, due to the utilized alignment board, the proposed approach can be seamlessly integrated within a stereo camera calibration scheme. Please note that within this work no assumption about the calibration of the deployed cameras is made. Thus, the problem of temporal alignment is solved for uncalibrated cameras.

The structure of this paper is as follows: Before introducing the overall temporal alignment approach in Section III, Section II discusses the proposed feature point localization scheme. Experimental results when employing the proposed framework with a number of unsynchronized IR/visible-light video sequences are presented in Section IV. Finally, our conclusions are given in Section V.

II. Feature point localization

Due to the different spectral sensitivity of IR and visible-light cameras, the construction of an alignment board which appears likewise in the visible-light and IR spectrum is not a trivial task. However, inspirations can be drawn from the field of stereoscopic IR/visible-light camera calibration where, just as in our case, a set of feature points has to be located in both modalities before computing the intrinsic and extrinsic camera parameters.

Given this fact a number of calibration devices have been proposed in the literature. For instance, Praskash et al. [10] advocate a heated chessboard as an appropriate calibration device. They argue that due to the different IR emissivity of black and white regions, it is possible to extract the corner points of the chessboard pattern in the visible-light and IR modality, respectively. Another strategy is chosen in [11], where a planar black/white checkerboard pattern, augmented by a set of resistors mounted in the centroids of each square, is used as a calibration device. In their approach, the corners of the black/white squares appear in the visible-light images whereas the energized resistors appear in the IR spectrum.

The alignment board chosen in this work involves the use of miniature light bulbs, equidistantly mounted on a planar calibration board [12]. This configuration is of special interest since, when turned on, heat and light are simultaneously emitted by the light bulbs causing the pattern to appear in both the visible-light and IR spectrum. This is demonstrated in Fig. 1, where the employed alignment board consisting of 81 light bulbs, arranged in a 9×9 matrix, is shown in the visible-light and IR spectrum. Please note that the depicted images were chosen arbitrarily from an IR/visible-light image sequence pair after successful temporal alignment.

Our feature point localization algorithm attempts to extract the sub-pixel positions of the miniature light bulbs along all video frames exhibiting the alignment board of Fig. 1. For this purpose, we start off by binarizing the individual video frames using the adaptive thresholding scheme of [13], separating the light bulb regions from the background. The result of the thresholding operation, employed on the IR and visible-light image of Fig. 1, respectively, is shown in Fig. 2. Please note that the wrongly extracted regions belonging to the watermark located in the upper right corner of the IR images (see Fig. 1) were excluded using a simple k-means clustering operation [14].

After the thresholding operation, the extracted light bulb regions may appear with irregular shapes in the binarized image and may no longer resemble the expected ellipsoidal radiation pattern. Thus, as a second step, the extracted light bulb regions need to be further processed. This is accomplished by fitting an ellipse to the boundary pixels of each region and by subsequently using the area of the computed ellipse as the new light bulb region. In our implementation we utilize the Canny edge detector [15] to compute the region boundaries. The ellipse fitting is performed by employing the algorithm of [16] which calculates the ellipses using a least-squares-based algorithm. The left-hand side of Fig. 3 shows the extracted ellipse for a single light bulb region from the IR image of Fig. 1.

A first estimate of the feature point positions is obtained by calculating the centers of mass of the refined light bulb regions within the original IR and visible-light images, respectively. However, due to measurement noise, in general, the estimated points do not correspond exactly to the "true" feature point locations. Thus, as a further step, we propose to utilize the

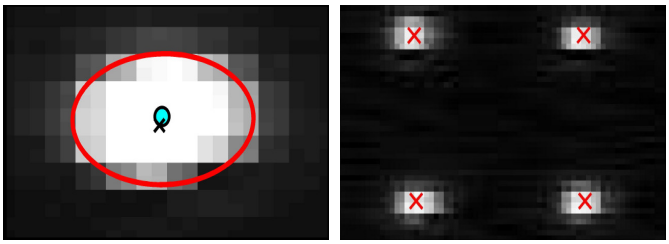

Fig. 3. Results of the feature point localization. (Left) Zoomed version of the IR image of Fig. 1 showing a single light bulb region with superimposed ellipse, and the obtained feature point before (circle) and after (cross) the minimization of eq. 1. (Right) Final feature point positions for a few selected light bulb regions.

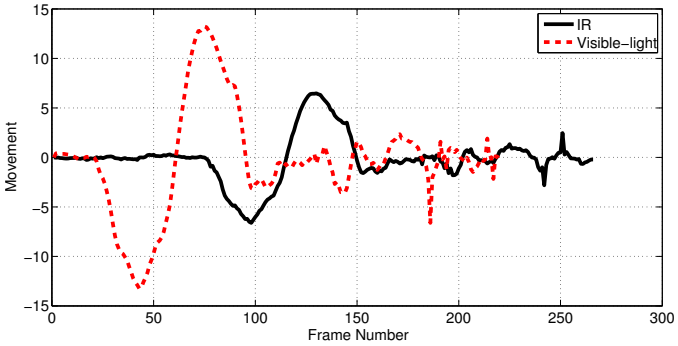

Fig. 4. Example of the vertical component of the speed of a single feature point along an IR video sequence (solid line) and a visible-light video sequence (dashed line).

available a-priori knowledge of the light bulb locations on the alignment board to improve the precision of the computed feature point locations.

For this purpose, let us introduce some concepts of projective geometry. In more detail, let the physical position of a feature point on the alignment board be represented by the homogeneous vector $\mathbf{x} = [x\ y\ 1]^T$ and its counterpart within an arbitrary image of an IR/visible-light image sequence pair as $\mathbf{x}' = [x'\ y'\ 1]^T$. Now, given a set of feature points $\mathbf{x}_i$ and a corresponding set of feature points $\mathbf{x}'_i$, there exists a projective transformation that takes each $\mathbf{x}_i$ to $\mathbf{x}'_i$. Thus, our goal is to find a 3×3 homography matrix $\mathbf{H}$ such that $\mathbf{H}\mathbf{x}_i = \mathbf{x}'_i$ for each i. However, due to the fact that the feature point locations obtained in the previous step were measured inexactly, there may not exist an homography which is able to map the feature point positions from the alignment board to the IR/visible-light images. A common solution to this problem is the usage of the Direct Linear Transformation (DLT) algorithm [7] which results in a first approximation of the underlying homography. However, this approximation is not optimal since it is obtained by employing the singular value decomposition which minimizes an algebraic distance measure that is not physically meaningful. Thus, we further refine the resultant homography by using the following cost function

$$\sum_i \left\| \mathbf{x}'_i - \mathbf{H}\mathbf{x}_i \right\|^2 , \qquad (1)$$

where $i = 1, \ldots, 81$ indexes the positions of the corresponding feature points. The final, refined homography is the one for which the geometric error given by eq. (1) is minimized using e.g. the iterative Levenberg-Marquardt algorithm [7].

The final feature point locations are obtained by applying the refined homography to the feature point positions within the alignment board coordinate system. The left-hand side of Fig. 3 shows the resulting feature point position for a single light bulb region before and after minimizing the cost function of eq. (1). The final feature point positions for a few selected light bulb regions of the IR image of Fig. 1 are depicted in the right-hand side of Fig. 3.

It will be shown in the next section that by means of these extracted feature point positions, the time-shift between two unsynchronized IR and visible-light video sequences can be successfully determined.

III. Temporal Alignment

Let $\mathbf{S}_V$ and $\mathbf{S}_I$ be two video sequences N_V and N_I long, recorded at the same frame rate by a visible-light and IR camera, respectively, exhibiting different poses of the alignment board of Fig. 1. Finding the temporal offset $\Delta\hat{t}$ between the two video sequences $\mathbf{S}_V$ and $\mathbf{S}_I$ is equivalent to maximizing a similarity measure $s(\cdot)$ over a set of potential temporal offset candidates Δt such that

$$\Delta\hat{t} = \arg\max_{\Delta t}\ s\big(\mathbf{S}_V, \mathbf{S}_I, \Delta t\big). \qquad (2)$$

Please note that in what follows, we start from the premise that the two video cameras are mounted side by side on a rail, hence, we assume that their position only differs horizontally and is identical otherwise.

The proposed temporal alignment approach starts off by performing translational movements of the alignment board in the downward and upward direction, respectively. This is followed by the extraction of the feature point positions in each frame of the IR and visible-light video sequence as elaborated in Section II. Based on the extracted feature point positions, we determine the vertical component of the speed of each feature point along the video sequence. This is accomplished by subtracting the y-coordinates of the feature point positions between two successive video frames. Fig. 4 shows an example of the vertical speed of a single feature point along an IR/visible-light video sequence pair. From the depicted curves the downward and upward swing of the alignment board, given by the negative and positive portions of the curves, respectively, can be seen. This is particularly apparent when looking at the dashed line, representing the visible-light video sequence. Please note that the scale difference between the amplitudes of the two curves is due to a mismatch in image resolution between the IR and visible-light video sequences.

Another way to look at the problem of temporal alignment is presented in Fig. 5. Here, the global movement of all 81 feature points is displayed as an image, with each line representing the overall vertical movement of a single feature point. In both images, brighter pixel values indicate the displacement of the alignment board in the upward direction whereas darker pixel values suggest a downward movement of the alignment

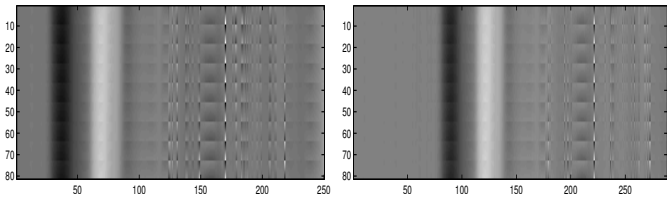

Fig. 5. Global movement of all 81 feature points within the visible-light (left) and IR video sequence (right). Each line represents the vertical movement of a single feature point. Bright pixel values indicate an upward movement whereas dark pixel values represent a downward movement of the alignment board.

board. Based on Fig. 5, the temporal offset we are looking for corresponds to the horizontal displacement between the two images such that the horizontal cross-correlation is maximized.

Let us put this observation now in a mathematical context. Given a temporal offset candidate Δt, the similarity between the visible-light sequence $\mathbf{S}_V$ and the IR sequence $\mathbf{S}_I$ is given by

$$s(\mathbf{S}_V, \mathbf{S}_I, \Delta t) = \frac{\sum_{m=1}^{M} \sum_{n \in \mathcal{N}} \mathbf{M}_V(m, n - \Delta t) \mathbf{M}_I(m, n)}{\sqrt{\sum_{m=1}^{M} \sum_{n \in \mathcal{N}} \big(\mathbf{M}_V(m, n - \Delta t)\big)^2 \sum_{k=1}^{K} \sum_{l \in \mathcal{N}} \big(\mathbf{M}_I(k, l)\big)^2}}, \quad (3)$$

where the matrices $\mathbf{M}_V(m, n)$ and $\mathbf{M}_I(m, n)$ express the movement of the m^{th} feature point between two consecutive visible-light and IR frames at time instant n, respectively, and $\mathcal{N} = \{\forall\ n \mid 1 \leq (n - \Delta t) \leq N_V \wedge 1 \leq n \leq N_I\}$. Please note that the similarity measure of eq. (3) is bounded to the interval $[-1, 1]$. The two video sequences are considered identical if the similarity measure is 1 and complementary to each other if the result is -1. A result of 0 implies that no similarities between the two sequences could be found. Finally, as expressed in eq. (2), the actual temporal offset $\Delta\hat{t}$ between the IR and visible-light video sequence is the one for which eq. 3 is maximized.

Fig. 6 shows the results of the temporal alignment for the two IR/visible-light video sequences corresponding to Fig. 5. It can be observed that the highest similarity (according to eq. (3)) is obtained for a temporal offset of 51 frames. This result corresponds well with Fig. 5 which, when evaluated subjectively, suggests a time-shift of approximately 50 frames between the two sequences.

IV. Results

In order to show the effectiveness of the proposed framework for the temporal alignment of IR/visible-light video sequence pairs, we performed experiments with five different, manually recorded data sets. Apart from the actual scene content (which may vary from sequence to sequence), each video exhibits different poses of the alignment board of Fig. 1. These poses include translational and rotational movements of the alignment board and were chosen in such a way that,

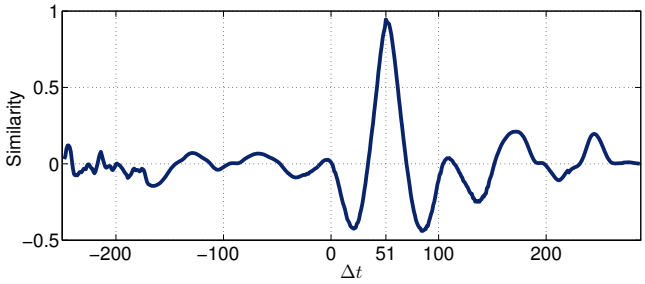

Fig. 6. Result of the temporal alignment for the two IR and visible-light video sequences corresponding to Fig. 5. The highest similarity (according to eq. (3)) between the two video sequences is obtained for a temporal offset Δt of 51 frames.

TABLE I

Results of the temporal offset estimation for five different IR/visible-light video sequence pairs.

	1st pair	2nd pair	3rd pair	4th pair	5th pair
Temporal Offset	54	55	67	51	69
Similarity	0.9881	0.9612	0.9805	0.9455	0.9869

both, temporal alignment and a future stereo calibration of the employed cameras can be performed simultaneously.

The IR video sequences were obtained by recording the analogue NTSC video output of a FLIR Prism DS camera, operating at a spectral range of 3.6 to 5 μm. In order to convert the analogue video stream to digital video, a Blackmagic Decklink HD Extreme 3D video capturing card was utilized. In accordance with the NTSC standard, the resultant video exhibits a resolution of 720×486 (which differs from the native resolution of the employed IR camera, 320×244). As for the visible-light video sequences a Panasonic HDC-TM700 camera was employed. The corresponding videos exhibit a resolution of 1920×1080. Both, IR and visible-light video sequences were recorded at a rate of 30 frames per second.

The estimated temporal offsets $\Delta\hat{t}$ for all five tested video sequences together with the corresponding similarity measures of eq. (3) are given in Table I. Note that the attained similarity is close to one for all five tested scenarios. This implies that after temporal alignment the movements of the alignment board are almost identical between the IR and visible-light video sequences. However, it is worth noting that the overall similarity measure depends on the performed movements with the alignment board. Thus, a lower similarity does not necessarily suggest a poor estimation of the temporal offset. Furthermore, by plotting the similarity measures over the whole set of temporal offset candidates (see Fig. 6) for each tested video pair, we observed that the curves always exhibit a single distinct peak at the position of the correct temporal offset, indicating a high robustness of our framework.

Finally, in order to qualitatively demonstrate the effectiveness of our proposed temporal alignment scheme, Fig. 7 shows the superposition of five frames from the 5^{th} IR/visible-light video sequence pair before and after temporal alignment. It

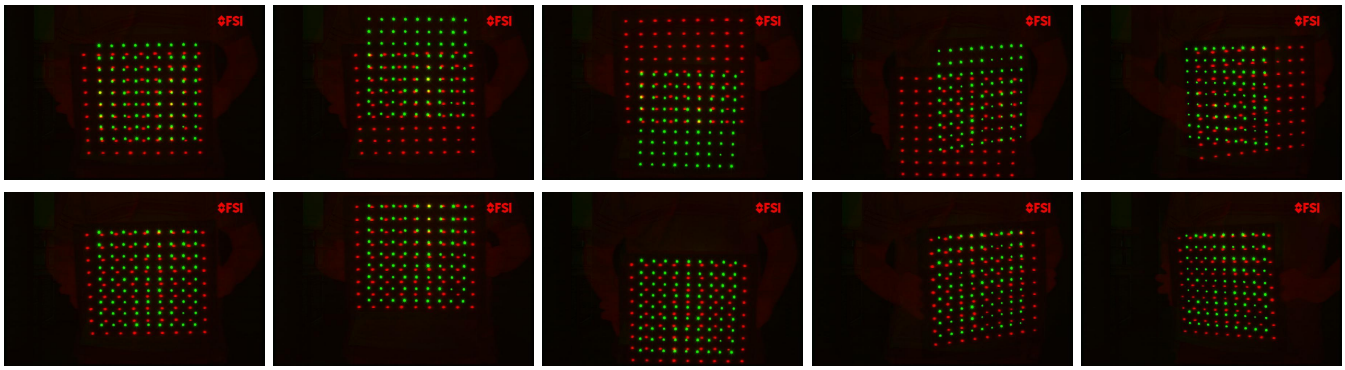

Fig. 7. Five representative frames (58, 88, 154, 223, 296) from the 5^{th} IR/visible-light video sequence pair before (top row) and after (bottom row) temporal alignment. For visualization purposes, corresponding frames of the two sequences were superimposed and occupy the red (IR images) and green (visible-light images) bands of the depicted RGB pseudo-color images.

can be noted that the unsynchronized video frames (top row) display a significant misalignment in time. This is particularly evident when observing the four IR video frames to the right which appear to lag considerably behind the visible-light frames. As for the synchronized video frames (bottom row), it can be seen that both IR and visible-light frames exhibit similar poses of the alignment board, thus, indicating the correct temporal alignment of the IR/visible-light video sequence pair.

V. Conclusion

In this work, a novel technique for the robust temporal alignment of IR and visible-light video sequences was introduced. Our method utilizes a planar alignment device equipped with miniature light bulbs to increase the number of corresponding feature points within the frames of a misaligned IR/visible-light video sequence pair. Thereby, the alignment process is turned more robust against the chronic lack of mutual scene characteristics, which represents a common source of problems when synchronizing video sequences originating from different spectral modalities.

The proposed processing chain first determines the exact light bulb positions in the individual frames of both IR and visible-light video sequences and computes the overall movement of the feature points along the video sequences. Subsequently, these movements are used to solve for the correct temporal offset.

We assessed our framework by performing experiments with five different, misaligned IR/visible-light sequence pairs. These tests suggested that the proposed temporal alignment technique is able to estimate the temporal offset with a very high confidence level.

Acknowledgements

The authors would like to thank the Brazilian Funding Agency CAPES (Pro-Defesa) for the financial support.

References

[1] Y. Caspi and M. Irani, "Spatio-temporal alignment of sequences," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 11, pp. 1409–1424, 2002.

[2] J. Yan and M. Pollefeys, "Video synchronization via space-time interest point distribution," in *Advanced Concepts for Intelligent Vision Systems*, 2004, pp. 501–504.

[3] Y. Caspi, D. Simakov, and M. Irani, "Feature-based sequence-to-sequence matching," *International Journal of Computer Vision*, vol. 68, no. 1, pp. 53–64, June 2006.

[4] L. Wolf and A. Zomet, "Wide baseline matching between unsynchronized video sequences," *International Journal of Computer Vision*, vol. 68, no. 1, pp. 43–52, June 2006.

[5] C. Lei and Y.-H. Yang, "Tri-focal tensor-based multiple video synchronization with subframe optimization," *IEEE Transactions on Image Processing*, vol. 15, no. 9, pp. 2473–2480, 2006.

[6] F.L.C. Padua, R.L. Carceroni, G.A.M.R. Santos, and K.N. Kutulakos, "Linear sequence-to-sequence alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 2, pp. 304–320, 2010.

[7] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2nd edition, 2004.

[8] Y. Ukrainitz and M. Irani, "Aligning sequences and actions by maximizing space-time correlations," in *Proceedings of the 9th European Conference on Computer Vison*, 2006, vol. 3953, pp. 538–550.

[9] M. Ushizaki, T. Okatani, and K. Deguchi, "Video synchronization based on co-occurrence of appearance changes in video sequences," in *Proceedings of the 18th International Conference on Pattern Recognition*, 2006, vol. 3, pp. 71–74.

[10] S. Prakash, P.Y. Lee, T. Caelli, and T. Raupach, "Robust thermal camera calibration and 3D mapping of object surface temperatures," in *Proceedings of the XXVIII SPIE Conference on Thermosense*, 2006, vol. 6205.

[11] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Infrared camera calibration for dense depth map construction," in *Proceedings of the 2011 Intelligent Vehicles Symposium*, June 2011, pp. 857 – 862.

[12] A. Ellmauthaler, Pagliari C.L. da Silva, E.A.B., and Gois J.N., "A novel iterative calibration approach for thermal infrared cameras," *submitted for publication in Proceedings of the 2013 IEEE International Conference on Image Processing*.

[13] C.D. Prakash and L.J. Karam, "Camera calibration using adaptive segmentation and ellipse fitting for localizing control points," in *Proceedings of the 2012 IEEE International Conference on Image Processing*, October 2012, pp. 341–344.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Verlag, New York, 2nd edition, 2008.

[15] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679 – 698, November 1986.

[16] A.W. Fitzgibbon, M. Pilu, and R.B. Fisher, "Direct least-squares fitting of ellipses," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 5, pp. 476 – 480, May 1999.