TWO STATISTICAL MEASURES OF SIMILARITY FOR OBJECT ASSOCIATION AND TRACKING IN COLOR IMAGE SEQUENCES

Hugh L. Kennedy

Technical Knockout Systems Pty. Ltd.
LPO Box 8150, ANU Acton, ACT 0200, Australia; hkennedy@tkosystems.com.au

ABSTRACT

Two statistical measures of similarity, for data association and tracking moving objects in sequences of color images, are derived and their performance is compared with normalized cross-correlation. Both methods use an F-distributed test statistic in a hypothesis test, which permits association thresholds to be set to give the desired (theoretical) false-association rate. One of the methods matches the performance of normalized cross-correlation, in the test data used, and is computationally less expensive.

Index Terms— Image motion analysis, Spatial filters, Image matching, Image classification, Tracking.

1. INTRODUCTION

Automatic processing in surveillance systems (e.g. [1]), for finding and following potential threats in image sequences or video streams, is typically composed of detection and tracking components. The detection component identifies regions of change, from one image (or frame) to the next [2,11]; while the tracking component fuses information from consecutive frames together over time. For analysis and design purposes, tracking may be factored into association and estimation processes. Association is the process whereby all detections due to a given target are identified; whereas estimation uses the associated detections to reduce position and velocity errors. This paper is primarily concerned with the problem of association.

Association and estimation are not independent processes, as good track estimates make it easier to find the correct detection; and correct association prevents false measurements, absent target detections, or detections due to other targets, from degrading track estimates. Tracking in high-resolution feeds from optical sensors is perhaps unique in that the target detection characteristics (size, shape, texture, and possibly color) are typically well defined in the digitized data matrix (i.e. the image). The effective utilization of this information greatly simplifies the association problem. This observation has, in part, motivated the development of track-before-detect algorithms [3], which operate on features directly in the discrete measurement space instead of extracted peaks. In principle, reliable association makes it possible to find the detection due to a given target anywhere in the measurement space, without a prior estimate of the target's state.

The problem of association is closely related to the problems of automatic classification [4], pattern recognition [5], image registration [6] and object detection [7]. The methods used to solve these problems vary widely; however, the use of templates and some measure of "similarity" [5,7,9], "distance" [4], "correlation", "alignment", "coincidence" or "correspondence" [6], is common. The technique of cross correlation and methods derived from it, such as normalized cross-correlation and normalized covariance [6,8], may be used to align a portion of an image (or signal) with a template. The translation which maximizes the similarity measure is taken to be the most likely. Other methods such as Sum of Squared Differences (SSD) [12,13] and Sum of Absolute Differences (SAD) [14] are also commonly used (these values are minimized, not maximized). The SAD method is generally favored in real-time image processing applications because its low computational complexity allows it to quickly process large volumes of data. A number of alternative methods have also been proposed [7,9]. Two new methods [10] are described below. Data sets of varying degrees of difficulty are given, and performance of the methods is discussed, relative to each other and to normalized cross-correlation.

2. PRELIMINARIES

Image data are represented using a three dimensional array where $I(i,j,k)$ is the intensity of the k th color at the pixel in the i th row and j th column. Consider two images I_1 and I_2 of equal size, with m rows, n columns and K colors, captured at times t and $t+\Delta t$ respectively. A moving object has been found in each image by a change detector. The detector has determined that object's centre is at $\left(i_1^{\max}, j_1^{\max}\right)$ and $\left(i_2^{\max}, j_2^{\max}\right)$ in images 1 and 2 respectively. The detector uses a rectangular box to completely enclose the object in each image, with the lower and upper limits in the vertical and horizontal dimensions, defined using i_1^{lwr}, i_1^{upr}, j_1^{lwr}, j_1^{upr}, i_2^{lwr}, i_2^{upr}, j_2^{lwr}, and j_2^{upr}. The portion of the image that is bound by the box (a detection) in each frame, D_1 and D_2, that are deemed to be due to the same target by the associator (described below), is retained and passed to the estimator (not described here). Detections in any given frame (I_1) are carried forward and used as templates during association when a later frame (I_2) is received. All previous detections in I_1 are correlated with all current detections in I_2. The term "correlated" is used here to collectively refer to all processes described below, which involve the alignment of the two detections by maximizing a Measure Of Similarity (MOS). The alignment process slides the template through all vertical and horizontal translations (a and b) that keep the centre of D_1 within the box defining D_2. The translation for which the MOS is maximized is selected, and the

maximum value retained. A faster (but less reliable) correlation process is to use only one translation, so that the centers of the detections coincide (not investigated here).

Even when there is only one detection in each image, association is not unambiguous, if the possibility of false alarms, and variable target visibility and variable target existence are considered. For this reason, a way of determining the (minimum) thresholds for the MOS is needed, preferably by theoretic means. Methods A and B described below, provide a statistical basis for setting the threshold, which may be set to give the desired false-association rate, when the null hypothesis is indeed true. When multiple detections are present a so-called "greedy" assignment algorithm is used, with the test statistic used as the MOS.

3. NORMALIZED CROSS-CORRELATION

Normalized cross-correlation (ρ_{12}) is preferable to (standard) non-normalized correlation (R_{12}), not only because ρ_{12} bounded between -1 and +1, which facilitates the arbitrary selection of a threshold (e.g. 0.9), but it also has been found to be less likely to yield false peaks than its non-normalized counterpart [8,9]. When processing non-zero-mean data such as images, the mean is usually first subtracted using

$$J(i,j,k) = I(i,j,k) - \frac{1}{N}\sum_{i'=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j'=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}} I(i',j',k), \tag{1}$$

where $N = \left(i^{\mathrm{upr}} - i^{\mathrm{lwr}} + 1\right)\left(j^{\mathrm{upr}} - j^{\mathrm{lwr}} + 1\right)$, i.e. the number of pixels within the analysis window, for a given color. The normalized cross-correlation coefficient, for multiple colors is, then computed using

$$\rho_{12}(a,b) = \frac{\sum_{i=i_1^{\mathrm{lwr}}}^{i_1^{\mathrm{upr}}}\sum_{j=j_1^{\mathrm{lwr}}}^{j_1^{\mathrm{upr}}}\sum_{k=1}^{K} J_1(i,j,k)J_2(i-a,j-b,k)}{\sqrt{\sum_{i=i_1^{\mathrm{lwr}}}^{i_1^{\mathrm{upr}}}\sum_{j=j_1^{\mathrm{lwr}}}^{j_1^{\mathrm{upr}}}\sum_{k=1}^{K} J_1(i,j,k)^2}\sqrt{\sum_{i=i_1^{\mathrm{lwr}}}^{i_1^{\mathrm{upr}}}\sum_{j=j_1^{\mathrm{lwr}}}^{j_1^{\mathrm{upr}}}\sum_{k=1}^{K} J_2(i-a,j-b,k)^2}}, \tag{2}$$

for all displacements (a,b) of interest (see previous section), e.g.

$$a = i_1^{\max} - i_2^{\mathrm{upr}} \ldots i_1^{\max} - i_2^{\mathrm{lwr}} \tag{3a}$$

and

$$b = j_1^{\max} - j_2^{\mathrm{upr}} \ldots j_1^{\max} - j_2^{\mathrm{lwr}}. \tag{3b}$$

The summations are truncated where the analysis window extends beyond the image boundaries.

4. METHOD A

A linear model is used to describe the appearance of the target from one frame to the next,

$$J_2(i,j,k) = \beta(k)J_1(i,j,k) + \varepsilon, \tag{4}$$

where the error term ε is Gaussian-distributed, zero-mean, white noise, i.e. $\varepsilon \sim \mathrm{N}\{0,\sigma^2\}$, and the indices are confined to the detection box. The color-dependent parameter β is used to model changes in illumination, reflectivity and sensor sensitivity; while the color-independent parameter σ^2 is used to model all other unknown causes of intensity fluctuation. Their Maximum Likelihood Estimates (MLEs) are computed using

$$\hat{\beta}(k) = \frac{1}{\alpha(k)}\sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}} J_1(i,j,k)J_2(i,j,k) \tag{5}$$

with

$$\alpha(k) = \sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}} J_1(i,j,k)J_1(i,j,k) \tag{6}$$

(a normalizing factor) and

$$\hat{\sigma}^2(k) = \frac{1}{N}\sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}}\left(J_2(i,j,k) - \hat{\beta}J_1(i,j,k)\right)^2, \tag{7}$$

although, the un-biased form of the variance is used in what follows, i.e. $\tilde{\sigma}^2 = N\hat{\sigma}^2/(N-1)$. From the theory of General Linear Models (GLMs), if

$$\alpha(k)\left(\hat{\beta}(k) - \beta(k)\right)^2 / \sigma^2 \sim \chi^2\{1\} \tag{8}$$

then

$$\frac{1}{\sigma^2}\sum_{k=1}^{K}\alpha(k)\left(\hat{\beta}(k) - \beta(k)\right)^2 \sim \chi^2\{K\}; \tag{9}$$

similarly, if

$$(N-1)\tilde{\sigma}(k)^2 / \sigma^2 \sim \chi^2(N-1) \tag{10}$$

then

$$\frac{(N-1)}{\sigma^2}\sum_{k=1}^{K}\tilde{\sigma}(k)^2 \sim \chi^2\{K(N-1)\}, \tag{11}$$

where $\chi^2\{v\}$ is the chi-squared distribution with v degrees of freedom.

Dividing (9) by (11) cancels σ^2 furthermore, dividing the numerator and the denominator by their respective degrees of freedom, yields a test statistic that is distributed according to Snedecor's F distribution, i.e.

$$Z_A = \sum_{k=1}^{K}\alpha(k)\left(\hat{\beta}(k) - \beta(k)\right)^2 \Big/ \sum_{k=1}^{K}\tilde{\sigma}(k)^2 \sim \mathrm{F}\{K, K(N-1)\}. \tag{12}$$

The null hypothesis is that $\beta(k) = 0$ for all k, i.e. that the two detections are orthogonal. If the null hypothesis is rejected then the two detections are declared to be similar. The Z_A statistic is evaluated at all feasible displacements by translating one of the image portions relative to the other, as done in (2), prior to least-squares fitting.

5. METHOD B

This method begins with the null hypothesis, that both detections contain only zero-mean Gaussian-distributed noise, i.e.

$$J_1(i,j,k) \sim \mathrm{N}\{0,\sigma^2\} \text{ and } J_2(i,j,k) \sim \mathrm{N}\{0,\sigma^2\}. \tag{13}$$

While this is a reasonable assumption for acoustic measurements [9], it is only reasonable for images after high-pass filtering, or mean subtraction, as in (1). Summing then squaring (13), and differencing then squaring, J_1 and J_2, yields

$$\left(J_1(i,j,k)+J_2(i,j,k)\right)^2 \Big/ 2\sigma^2 \sim \chi^2\{1\} \quad (14)$$

and

$$\left(J_1(i,j,k)-J_2(i,j,k)\right)^2 \Big/ 2\sigma^2 \sim \chi^2\{1\}, \quad (15)$$

respectively. From the reproductive property of chi-squared variables, summing these quantities over the analysis window, and over all colors, then dividing the former by that latter, cancels σ^2 and results in the following F-distributed test statistic:

$$Z_B = \frac{\sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}}\sum_{k=1}^{K}\left(J_1(i,j,k)+J_2(i,j,k)\right)^2}{\sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}}\sum_{k=1}^{K}\left(J_1(i,j,k)-J_2(i,j,k)\right)^2} \sim \mathrm{F}\{NK,NK\}. \quad (16)$$

Note that Z_B is equivalent to Z_M in [9], when $K=1$ in (16) and $M=2$ in [9]. The test statistic Z_B is computed at all candidate displacements using

$$Z_B(a,b) = \frac{\sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}}\sum_{k=1}^{K}\left(J_1(i,j,k)+J_2(i-a,j-b,k)\right)^2}{\sum_{i=i^{\mathrm{lwr}}}^{i^{\mathrm{upr}}}\sum_{j=j^{\mathrm{lwr}}}^{j^{\mathrm{upr}}}\sum_{k=1}^{K}\left(J_1(i,j,k)-J_2(i-a,j-b,k)\right)^2}. \quad (17)$$

This test statistic is evaluated and the zero-mean hypothesis tested. If the test statistic is significantly larger than would be expected, under the null hypothesis, then the null hypothesis is rejected. This test statistic is powerful and selective, as only similar and well-aligned objects fail the test when the size is set low. This is due to the fact that the denominator approaches zero for perfectly aligned images (making Z_B very large) while it remains large and of similar magnitude to the numerator when they are misaligned (keeping Z_B close to unity). Intense detections are also highly likely to result in large Z_B values because the sum in the numerator is large.

While methods A and B both provide a theoretical basis for setting the association threshold – to give the desired test size – empirical tuning is still required to give an acceptable false association rate in real environments (where the null hypotheses are never entirely true) and a satisfactory probability of correct association for real targets of interest.

6. RESULTS

The methods were compared using an approximately equal number of easy (textured background) and hard (cluttered background) data sets, giving rise to a total of 124 detections for analysis. All images were captured and processed using 8 bit (RGB) color. No thresholds were used to process the data; as a consequence, the most likely association hypothesis was always chosen, regardless of its magnitude. Examples of I_2 are shown in Figure 1 and Figure 3. A change detector based on the method described in [2] was used to identify the detections. For large displacements between frames, this detector usually identifies two detections – one for the departing object and one for the arriving object. The detection boxes are drawn in white. Every detection in I_1 was correlated with every detection in I_2 using the methods described above, and the most likely assignment hypotheses selected. This was done independently, on a pairwise basis, and no attempt was made to produce a global multi-target assignment solution. The assignment success rate (see Table 1) was determined by visual inspection. Only detections containing foreground objects in I_1 were considered; detections on empty background or false alarms were not considered. Detections in I_1 with no detection in I_2, due to detector error, were also disregarded. Sample results of association computations are shown in Figures 2 and 4.

TABLE 1. Assignment success rate.

	Easy. Mean Sub.	Easy. Mean Not Sub.	Hard. Mean Not Sub.
Z_A	56%	80%	33%
Z_B	59%	84%	56%
ρ_{12}	59%	84%	54%

FIGURE 1. Easy example image.

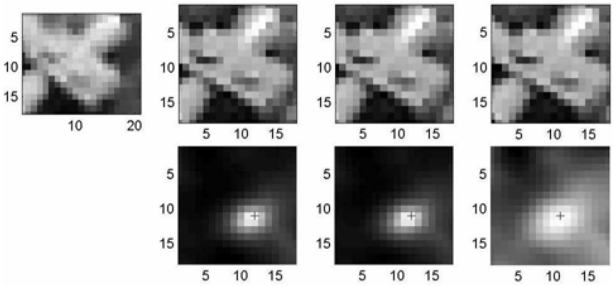

FIGURE 2. Easy example results.

FIGURE 3. Hard example image.

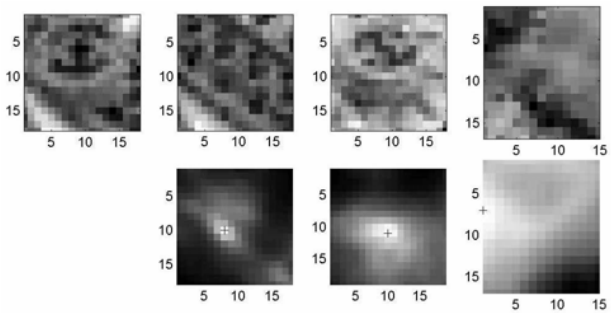

FIGURE 4. Hard example results.

7. DISCUSSION

Subtracting the mean, prior to correlation, degrades the performance of all methods significantly, it is therefore not recommended. However it was noted, for all methods, that mean subtraction does increase the difference between the MOS for the worst detection and the best detection which potentially makes it easier to set thresholds to optimize the association behavior. The performance of the Z_B and ρ_{12} methods were similar; both outperformed the Z_A method. The Z_A and Z_B methods both yield sharper peaks and flatter backgrounds than the ρ_{12} method, in the correlation matrices (lower row of Figures 2 and 4) although this does not necessarily mean improved performance. The top left image in these figures is a given detection in I_1; the remaining images are the portion of I_2 which the maximize the MOS. The three right-most columns are for Z_A, Z_B and ρ_{12}, respectively. In Figure 4 only Z_B yields the correct result; while Z_A associates with the background – a common problem for all methods. Judging from the (model) implementation used to process these data, Z_A is approximately three times slower than Z_B; while Z_B is approximately 30% faster than ρ_{12}. The latter observation is expected, as (17) requires fewer floating-point operations than (2). Both Z_B and ρ_{12} can be computed very efficiently using circular buffers to compute moving averages (not implemented here).

8. CONCLUSION

Methods A and B are based on hypothesis tests, with the test statistic used as a measure of similarity (MOS). Like normalized cross-correlation, method B provides an effective means of associating similar objects. The low computational complexity of method B makes it an attractive alternative to normalized cross-correlation in real-time applications.

9. REFERENCES

[1] S. Muller-Schneiders, T. Jager, H.S. Loos and W. Niem, "Performance evaluation of a real time video surveillance system", in *Proc. 2nd Joint IEEE Int. Workshop on VS-PETS,* Beijing, pp. 137-143, Oct. 2005.

[2] T. Aach, A. Kaup and R. Mester, "Statistical model-based change detection in moving video", *Signal Processing*, vol. 31, no. 2, pp. 165-180, Mar. 1993.

[3] Wei Zhang, Mingyu Cong and Liping Wang, "Algorithms for optical weak small targets detection and tracking: review", in *Proc. 2003 Int. Conf. on Neural Networks and Signal Process.*, Nanjing, China, vol. 1, pp. 643-647, Dec. 2003.

[4] D. Toth and T. Aach, "Improved minimum distance classification with Gaussian outlier detection for industrial inspection", in *Proc. 11th Int. Conf. on Image Anal. Process.*, Palermo, Italy, pp. 584-588 Sept. 2001.

[5] A.K. Jain, R.P.W. Duin, Jianchang Mao, "Statistical pattern recognition: a review", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 4-37, Jan. 2000.

[6] D.N. Bhat and S.K. Nayar, "Ordinal measures for image correspondence", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 4, pp. 415-423, Apr. 1998.

[7] J. Ben-Arie and K.R. Rao, "Template recognition based on expansion matching with neural lattice implementation", in *Vision Models for Target Detection and Recognition*, E. Peli (Ed.), World Scientific, Singapore, pp. 284-310, 1995.

[8] F. Viola and W.F. Walker, "A comparison of the performance of time-delay estimators in medical ultrasound", *IEEE Trans. Ultra., Ferro. Freq. Contr.*, vol. 50, no. 4, pp. 392–401, Apr. 2003.

[9] H.L. Kennedy, "A new statistical measure of signal similarity", *Proc. International Symposium on Information, Decision and Control 2007*, pp. 112-117, Adelaide, Australia, Feb. 2007.

[10] Patent Pending, Provisional Application No. 2006901479, "Signal Analysis Methods", Mar. 2006.

[11] R.J. Radke, S. Andra, O. Al-Kofahi and B. Roysam, "Image change detection algorithms: a systematic survey", *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294-307, Mar. 2005.

[12] A. Giachetti, M. Campani and V. Torre, "The use of optical flow for road navigation," *IEEE Trans. Robotics and Automation*, vol. 14, no. 1, pp. 34-48, Feb 1998.

[13] Teahyung Lee and D. Anderson, "Performance analysis of a correlation-based optical flow algorithm under noisy environments," *Proc. 2006 IEEE Int. Symp. on Circuits and Systems (ISCAS 2006),* pp. 4699-4702, May 2006.

[14] H. Inoue, T. Tachikawa and M. Inaba, "Robot vision system with a correlation chip for real-time tracking, optical flow and depth map generation," *Proc. 1992 IEEE Int. Conf. on Robotics and Automation*, vol. 2, pp. 1621-1626, May 1992.