June 1996

Adaptive Correlation Tracking of Targets with Changing Scale

Ulf M. Cahn von Seelen  
*University of Pennsylvania*

Ruzena Bajcsy  
*University of Pennsylvania*

Follow this and additional works at: [http://repository.upenn.edu/cis_reports](http://repository.upenn.edu/cis_reports)

**Recommended Citation**  
[http://repository.upenn.edu/cis_reports/234](http://repository.upenn.edu/cis_reports/234)


This paper is posted at ScholarlyCommons. [http://repository.upenn.edu/cis_reports/234](http://repository.upenn.edu/cis_reports/234)
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).
Adaptive Correlation Tracking of Targets with Changing Scale

Abstract
Algorithms for tracking targets imaged through a zoom lens must accommodate changes in the magnification of the target. This requirement poses particular problems for correlation techniques, which usually are not invariant to scale changes. An adaptive correlation method has been developed that selectively updates the correlation template in response to scale changes in an image sequence. The algorithm estimates a subset of the parameters of the affine transformation between the template and the matched image patch and updates the template only when the scaling exceeds given bounds. The selective template update enables correlation to track targets at varying scale while decreasing the risk of template drift.

Comments
Adaptive Correlation Tracking of Targets with Changing Scale

MS-CIS-96-22
GRASP LAB 405

Ulf M. Cahn von Seelen
Ruzena Bajcsy

University of Pennsylvania
School of Engineering and Applied Science
Computer and Information Science Department
Philadelphia, PA 19104-6389

1996
Adaptive correlation tracking of targets with changing scale

Cahn von Seelen, Ulf M.
Bajcsy, Ruzena
GRASP Laboratory
Department of Computer and Information Science
University of Pennsylvania

June 1996

Abstract

Algorithms for tracking targets imaged through a zoom lens must accommodate changes in the magnification of the target. This requirement poses particular problems for correlation techniques, which usually are not invariant to scale changes. An adaptive correlation method has been developed that selectively updates the correlation template in response to scale changes in an image sequence. The algorithm estimates a subset of the parameters of the affine transformation between the template and the matched image patch and updates the template only when the scaling exceeds given bounds. The selective template update enables the correlation to track targets at varying scale while decreasing the risk of template drift.

1 Introduction

The overall objective of the project RSTA on the Move [Davis et al., 1996] is to equip a vehicle with an active camera system that can stabilize the image stream, detect a moving target, track it, and with a zoom lens magnify details of the target that may yield important reconnaissance information. The University of Maryland develops image stabilization algorithms that aim to maintain, ideally, a stable scene background. The stabilized images are the input to a module from the University of Rochester and the University of Maryland that detects moving targets. NIST and the University of Pennsylvania provide algorithms that track a detected target and control the pan, tilt, and zoom of the active camera system.

The work at the University of Pennsylvania focuses on tracking a target with a variable focal length lens. The ultimate goal is to control the focal length of the active vision system
such that the detected target is imaged with maximum resolution. At the same time, the
magnification must not be so high that it prevents the tracking system from maintaining
acquisition of the target.

This paper presents a tracking algorithm for targets that are imaged at varying scale.
Since the target is provided to the tracker in the form of a small region of interest, correlation
is the natural choice for tracking it over time.

Traditional correlation algorithms are invariant to target translation, but they assume
that the target does not change its appearance, i.e., its orientation (2-D and 3-D) and size.
However, scale changes are the main effect when a target of unchanging appearance is imaged
with a zoom lens and can therefore cause the tracker to lose acquisition. To maintain a lock
on the target, the underlying correlation algorithm must be made robust to the zoom-induced
scale changes.

Section 2 reviews the classical correlation techniques and a variety of approaches to
make them invariant to 2-D rotation, scaling, or even general transformations. Out of these
approaches, a new method is developed that allows the correlation process to adapt to
scale changes. The basis for this adaptive correlation method is an algorithm to estimate
scale change. It is introduced and experimentally validated in section 3. Section 4 reports on
experiments with the adaptive correlation technique on an image sequence containing camera
pan and zoom. Section 5 outlines how the algorithm can be improved and incorporated into
RSTA’s active vision platform.

2 Correlation

Template matching/correlation/matched spatial filtering for image registration is an idea
that goes back to the beginning of image processing and computer vision (see [Duda and
Hart, 1973] for a historical perspective). It computes the similarity of patches \( P \) in an image
\( I(x, y) \) to a template \( T(x, y) \). Locating a template in the image means finding the patch that
has the highest similarity with the template.

From minimizing the \textit{sum of squared differences}

\[
E(i, j) = \sum_{x,y \in P} (T(x, y) - I(x - i, y - j))^2
\]

and the assumption that the variation in the image energy is small, one obtains the classic
criterion of maximizing the \textit{cross-correlation}:

\[
R(i, j) = \sum_{x, y \in P} T(x, y) \cdot I(x - i, y - j)
\]

To be robust to local variations in the image energy, the \textit{normalized cross-correlation} is used:

\[
N(i, j) = \frac{\sum_{x,y \in P} T(x, y) \cdot I(x - i, y - j)}{\sqrt{\sum_{x,y \in P} T(x, y)^2} \cdot \sqrt{\sum_{x,y \in P} I(x - i, y - j)^2}}
\]
An alternative criterion is the correlation coefficient:

\[
C(i, j) = \frac{\text{Cov}(T, I)}{\sqrt{\text{Var}(T) \cdot \text{Var}(I)}} = \frac{\sum_{x,y \in \mathcal{P}} (T(x, y) - \bar{T})(I(x - i, y - j) - \bar{I})}{\sqrt{\sum_{x,y \in \mathcal{P}} (T(x, y) - \bar{T})^2} \cdot \sqrt{\sum_{x,y \in \mathcal{P}} (I(x - i, y - j) - \bar{I})^2}
\]

It is equivalent to the normalized cross-correlation of the mean-corrected template and image and is invariant to offsets in the absolute gray level and differences in the contrast of template and image. Both \(N(i, j)\) and \(C(i, j)\) are bounded by \(-1 \leq N(i, j), C(i, j) \leq +1\) so that the correlation result can be interpreted as a confidence value relative to the perfect match result of +1.

A thorough comparison of various correlation-type registration algorithms can be found in [Aschwanden and Guggenbühl, 1992]. The study evaluates 19 algorithms by varying template size and image degradations (decreasing luminance, Gaussian noise, salt-and-pepper noise, blurring, and magnification).

Correlation is translation-invariant but requires that template and image have the same scale and orientation. If this is not the case, the correlation value deteriorates, even if the unscaled and unrotated template is identical with the image. In [Mostafavi and Smith, 1978a; Mostafavi and Smith, 1978b] it is shown that in the case of an affine transformation between image and reference, there are template sizes and template shapes that minimize the probability of a mismatch as well as the registration error. The more severe the distortion is, the smaller the optimal template size becomes.

Many approaches have been suggested to overcome the scale and orientation dependence of correlation [Kumar, 1992]. Some methods perform the correlation after transforming the image and the template into a domain with different invariances. Examples are the Fourier-Mellin transform [Casasent and Psaltis, 1976], the log-polar transform [Messner and Szu, 1985], and the conformally mapped Wigner distribution [Jacobson and Wechsler, 1984]. Other approaches use only those features in the template that are invariant to scaling and rotation (e.g., circular harmonic components [Hsu et al., 1982]).

Since our work ultimately aims at a real-time implementation, we focus on correlation methods that work in the spatial domain, thus avoiding computationally expensive transformations, and make use of all the available template information. A variety of approaches can be found in the pattern recognition literature.

One way of achieving robustness to scaling and rotation is to provide templates at different scales and orientations. The real-time sum-of-squared-differences tracker of [Inoue, 1996], e.g., matches the image against precomputed rotated templates and handles scaling by subsampling the input image in hardware. The search space is kept small by using templates close to the one matched in the last frame. In [Yoshimura and Kanade, 1994], a basis set of eigenimages, rather than the rotated templates themselves, is used in the correlation to reduce the numbers of templates.

A higher-dimensional template space also underlies the work of [Zetzsche and Caelli, 1989; Caelli and Liu, 1988]. The authors represent patterns as sets of outputs of rotated
and logarithmically scaled Gaussian bandpass filters. Pattern recognition involves a 4-D correlation of the filtered image with the filter output sets of the pattern and is expensive to compute.

An explicit model of the transformation between image and reference is computed in the optimization procedure of [Grun and Baltsavias, 1985]. The model includes geometric and radiometric parameters. In [Lo and Gerson, 1979], the authors prove that the optimal template size decreases with increasing scaling between image and reference and estimate the parameters of an affine transformation by matching several smaller image patches. The resulting transformation is then used to update the reference position.

In the context of stereo matching, the work of [Remagnino et al., 1994] performs correlation between the images by iteratively minimizing a zero-mean sum-of-squared-differences functional. In order to correct for the distortion caused by the different viewpoints of the cameras, the functional includes not only translation terms but also parameters for an affine warp of one of the correlated image patches.

The pattern recognition scenario provides few constraints on the similarity between the template and the encountered image. However, in a tracking application such as required for the RSTA project, a good initial match between template and image has already been found, and the correlation process has to cope only with incremental changes in the appearance of the tracked object. A full search over the space of all template variations becomes unnecessary; instead, the template can be modified stepwise to keep pace with the appearance changes [Jackson, 1981; Tam et al., 1990; Montera et al., 1994; Parry et al., 1995]. Adaptive correlation is therefore our choice for the task of tracking with varying focal length.

There are two possibilities for updating the template. It can be warped/resampled, or it can be replaced by the matched image patch. Warping the template has the advantage that it does not cause template drift (see below) because no new image information is incorporated. On the other hand, the lack of new information causes the template to become blurred when the image is expanding due to increasing focal length. Warping also requires precise estimation of the transformation parameters. The simpler solution is to replace the template by its match in the image.

The main problem of template replacement is correlator walk-off due to the accumulation of matching errors and the inclusion of background pixels [Montera et al., 1994]. In consequence, the template can begin to drift away from the target if it is updated too frequently. Section 4.3 contains a detailed explanation of this phenomenon.

Several ways to reduce template drift have been proposed. In [Jackson, 1981], the template is only updated when the correlation score falls below a fixed threshold. In infrared imagery, thresholding the image provides target boundaries on which the new template can be recentered [Montera et al., 1994; Parry et al., 1995].

Our strategy for updating the template is based on measuring the scaling between the template and its match in the current frame. Only when the scaling becomes large enough
to impair the correlation, the template is updated. The next section introduces a differential method for estimating the translation and scaling between two images.

3 Estimating image scaling

3.1 Principle

If the target performs only translatory motions between frames (relative to the camera), its images are related by an affine transformation containing only a translation term and a scaling term:

\[ I(x, y, t) = I(x - (a + cx), y - (b + cy), t - 1) \]

If the scaling of the target changes only a little between frames, correlation will locate the template in the image with little error. Since the correlation process in this way corrects for most of the translation, the remaining translation (due to the single-pixel accuracy of the correlation) is small, and so is (by assumption) the scaling. The parameters can therefore be estimated by fitting an affine flow field to the template and the registered image patch using a differential approach [Bergen et al., 1990]: find parameters \( a, b, c \) such that

\[ \sum_{x,y \in P} (I(x, y, t) - I(x - a - cx, y - b - cy, t - 1))^2 = \min. \]

Since the motion is small, we can approximate \( I \) by the linear terms of its Taylor series expansion:

\[ I(x - a - cx, y - b - cy, t - 1) \approx I(x, y, t) - (a + cx)I_x(x, y, t) - (b + cy)I_y(x, y, t) - I_t(x, y, t) \]

The error to be minimized thus becomes:

\[ \sum_{x,y \in P} (I_t + (a + bx)I_x + (b + cy)I_y)^2 = \min. \]

Setting the partial derivatives of the error to zero leads to a system of three linear equations in the desired parameters \( a, b, c \):

\[
\begin{pmatrix}
\sum \frac{I_x^2}{I_x I_y} \\
\sum \frac{I_x I_y}{I_x^2} \\
\sum \frac{(x I_x + y I_y)^2}{I_x} \\
\end{pmatrix}
\begin{pmatrix}
\sum I_x I_t \\
\sum I_y I_t \\
\sum (x I_x + y I_y) I_t \\
\end{pmatrix}
\cdot
\begin{pmatrix}
a \\
b \\
c \\
\end{pmatrix}
=
\begin{pmatrix}
\sum I_x I_t \\
\sum I_y I_t \\
\sum (x I_x + y I_y) I_t \\
\end{pmatrix}
\]

The derivative computations and the summations take place over the template and the registered image patch, i.e., over a relatively small area. The whole algorithm therefore adds only little computational costs. It has been implemented, together with a correlation tracker, in OBVIUS, a Common Lisp-based image processing environment [Heeger et al., 1994].
3.2 Experimental validation

We examined the performance of the scale estimation algorithm on a zoom sequence of a toy tank (scale 1:35) at a distance of approximately 210cm from the sensor plane. The sequence was taken with a Sony XC77-RR b&w CCD camera and a motorized Fujinon H10xt1E-MPX31 zoom lens, 1:1.9/11-110mm. Camera and lens are mounted on a binocular camera platform, which is part of the PennEyes binocular active vision system [Madden and Cahn von Seelen, 1995]. Camera pan and lens zoom, focus, and aperture are under closed-loop control by a Delta Tau PMAC motion controller. In particular, a potentiometer provides absolute feedback about the focal length, with a nominal resolution of approximately 40000 encoder counts over the zoom range. The full range is swept in about 6 seconds. The actual magnification achieved over the full range is approximately 9.3. The pan axis is equipped with an optical quadrature encoder that provides position feedback with a nominal resolution of 0.006" (not taking into account the 5-bit sub-count interpolation). The axis can reach a peak velocity of 1000°/s at accelerations of 12000°/s².

For the zoom sequence, neither the camera nor the tank were moved. Only the focal length was increased over the span of 38 images from an estimated 13mm to an estimated 82mm, in steps of equal number of zoom encoder counts (which translate to focal length approximately in an inversely quadratic fashion). The actually measured magnification between first and last frame is approximately 6.0. The focus was adjusted manually for each frame.

Templates of size 16 × 16 and 32 × 32 were tracked through the sequence with a search radius of 8 and 16 pixels, respectively. Normalized cross-correlation was used, and the template was replaced with the best match after each frame. In this way a correspondence between templates on different scales was defined. Two starting locations were examined, the GRASP logo and the tank driver (figure 1).

When the magnification between template and current frame becomes too large, the scale estimation starts to fail, typically by returning a value that is lower than the cumulative magnifications computed for the previous frames. To determine the critical scaling value, each template in each frame served as the starting point of a tracking run through the rest of the sequence, and the last reliable estimated value was recorded.

We found that while the scale estimation algorithm underestimates the true scaling, it performs consistently and works up to magnification changes of 20%. This number should be compared to the highest change rate between successive images that can theoretically be achieved with the zoom lens if the images are taken at 30Hz. It is $10^{1/(6.0-30)} = 1.013$, i.e., 1.3%, and therefore lies well within the operating range of the scale estimation algorithm. Target motions in depth, of course, are theoretically unconstrained and can cause the algorithm to fail if they become too large. However, the scaling effects of the motion of a distant target are usually much smaller than the magnification changes induced by zooming.
Experiments

4.1 Setup

We tested the adaptive template update algorithm on a sequence of 51 images with mixed pan and zoom. The target is a stationary toy tank at approximately 227 cm distance from the sensor plane. The imaging apparatus and the correlation tracker implementation is the same as that employed for the experiments in section 3.2.

Frames 0 to 20 of the sequence consist of a slow panning motion of the camera, at a rate of 0.25 pixels/frame and a constant estimated focal length of 12.9 mm. From frame 20 to 30, the pan rate is decreased to 75% (in terms of the absolute angular displacement per frame) while the estimated focal length is doubled,\(^1\) in steps that are approximately equal in terms of magnification change. Frames 30 to 50 consist of another slow panning motion, again at a rate of 0.25 pixels/frame and a constant estimated focal length of 25.9 mm. Figure 2 shows selected frames from the image sequence.

4.2 Procedure

We examined the tracking behavior for two different template sizes: 32 x 32 with a search radius of 8, 16, and 32 pixels; and 16 x 16 with a search radius of 8 and 16 pixels. Normalized cross-correlation was used. Two different starting locations in the image were examined, the Grasp logo and the tank driver.

To establish an "ideal" template correspondence over the frames of the sequence, the correlation algorithm was run without template update during the pan segments. During

\(^1\)The actual magnification has been measured to be 1.8.
Figure 2: Frames 0, 20, 30, and 50 of the test sequence. Slow panning takes place during the whole sequence whereas the focal length is changed only from frame 20 to 30.
the zoom segment, the template was updated after each frame. Figure 3 shows the templates at the two starting locations in the first frame and at their final positions in the last frame of the sequence.

Figure 3: Position of test templates in the first and last frame (images enlarged)

The ideal template sequence provides a baseline against which the performance of various template update strategies can be evaluated. As a measure for the tracking error incurred by a strategy in each frame, we use the Euclidean distance between the template location computed by the particular strategy and the ideal template location. Three template update strategies were examined:

**Static**: Never update the template.

**Dynamic**: Update the template after each frame.

**Adaptive**: Update the template when the scale change exceeds a certain threshold.

For the update, the template is replaced by the best match in the current frame. A threshold value of 1.04 was used for the adaptive strategy.

### 4.3 Results

Figure 4 shows the performance of the various template update strategies for the different starting locations and template sizes. The curves for the different search radii are overlaid in each of the graphs because for the dynamic and the adaptive strategy they happen to be identical across radii. Only the static strategy shows an effect of the radius of the search (except for the "driver" template of size $32 \times 32$): as the radius increases, the error curves grow more steeply and reach higher values.
Figure 4: Performance of template update strategies for different template sizes and starting locations. Dashed lines = dynamic strategy, dotted lines = static strategy, solid lines = adaptive strategy.
The reason for this trend lies in the increased availability of local correlation maxima as the search radius increases. Initially, the static strategy maintains zero error because it never updates the template and therefore is incapable of causing template drift. During the zoom phase, however, the match of the template with the dilating image becomes increasingly worse, and the tracker latches on to one of the many other local maxima. (Apparently, the region around the logo location provides more opportunities for such diversions than the uniform background around the driver location because the logo template wanders off farther). The dynamic and the adaptive strategy are not affected by the search radius because they maintain a template that leads to high correlation values, which usually dwarf all other local maxima in the search window.

While the static strategy does not cause template drift but fails during zooming, the dynamic strategy shows the opposite behavior. It handles the zooming well but is susceptible to drift. This is apparent from the steady increase of the error over the course of the sequence.

The template drift is elicited by the slow panning motion in the test sequence. The sub-pixel motion causes small matching errors between a template taken from one frame and an image patch taken from the next. If the template is updated after each frame, the matching error is absorbed as a sub-pixel shift in the new template. After several frames, the series of small shifts in a common direction has added up to whole-pixel shifts of the template.

The adaptive strategy combines the good properties of the static and the dynamic strategies by limiting the template update to the zoom phase. During the pan segments, it does not update the template and therefore does not cause drift. The performance of this strategy is a function of the scale change threshold that triggers a template update. A good value corresponds to an increase (or decrease) in magnification that is likely to throw off the correlation. Determining such a value will remain an empirical issue because the likelihood of a mismatch depends on the template and the scene background.

In the experiments shown above, any threshold between 1.01 and 1.04 could have been used. As a matter of fact, a value of 1.03 or less would have led to the ideal template sequence (and thus to zero error) because the scale change between frames in the zoom segment is always greater than 1.03. We chose a threshold of 1.04 to demonstrate that it is sufficient to update the template only after the cumulative magnification over a few frames has exceeded a critical value.

5 Conclusions

We have presented a new technique to selectively adapt correlation templates in response to scale changes in an image sequence. The technique allows the correlation process to track targets at varying scale while minimizing the risk of having the template drift off the target. The key to its success is the adaptive template update strategy that triggers an update only when the magnification between the template and the current frame exceeds a threshold.
The fact that the estimation of the scaling puts an additional computational burden on
the tracking algorithm raises the question why the correlation score is not used as the update
criterion instead. For many applications this may be a viable solution and has indeed been
done (e.g., [Jackson, 1981]). However, some correlation measures do not provide an absolute
indicator of the match quality, e.g., unnormalized cross-correlation, and other ways must be
used to judge when the template ceases to yield reliable matches.

Estimating scaling isolates the single cause we want to react to from the many other
potential factors that can impair the match. It allows the algorithm to watch specifically for
scaling changes while preserving the template through temporary disturbances of other kinds,
e.g., partial occlusions. In addition, the translation parameters that are being estimated
along with the scaling can be put to use by interpolating the matching patch and replacing
the template with sub-pixel accuracy, which further reduces template drift.

In order to use the affine parameters for sub-pixel matching, the scale estimation algo-
rithm needs to be improved to recover the transformation parameters more accurately. Work
is also under way to implement a real-time version of the algorithm on our DSP-based image
processing system.

Acknowledgments

This research was supported by ARPA Grants DAAH04-93-G-0419 and N00014-92-J-1647;
Army Grants DAAH04-96-1-0007 and DAAL03-89-C-0031; and NSF Grants CDA88-22719,
CDA91-21973, IRI93-03980, IRI93-07126, and SBR89-20230.

References

ults from a comparative study on correlation-type registration algorithms. In W. Förstner
and St. Ruwiedel, editors, Robust Computer Vision, pages 268–289. Wichmann, Karlsruhe,
Germany, 1992.

Computing two motions from three frames. In 3rd International Conference on Computer

[Caelli and Liu, 1988] Terry M. Caelli and Zhi-Qiang Liu. On the minimum number of
templates required for shift, rotation, and size invariant pattern recognition. Pattern

[Casasent and Psaltis, 1976] David Casasent and Demetri Psaltis. Position, rotation, and


