
Masakatsu Kouogi  Takeshi Kurata  Katsuhiko Sakane
National Institute of Advanced Industrial Science and Technology (AIST)
1–1–1 Umezono, Tsukuba Chuo Dai 2, Tsukuba, Ibaraki, 305–8568, JAPAN
Phone: +81–298–61–2264, E-mail: m.kouogi@aist.go.jp

Abstract
In this paper, we describe an improved method of personal positioning and orientation using image registration between input video frames and panoramic images captured beforehand. In our previous work, we proposed the method of image registration with an affine transform. However, the affine transform is generally not capable of image registration between a frame and a panorama. We improved the previous method so that it can estimate projective transform parameters without severely increasing computational cost. We also improved the method to be robust with respect to lighting changes by using the weighted sum of absolute difference of both brightness and its gradient between images. We confirmed that this improved method could estimate image registration parameters under conditions that hindered the previous method. Its computational cost increased by only 10–20% and its software implementation was capable of real-time processing.

1 Introduction
An essential function of a wearable computer is to find the user’s location and orientation (context) relative to the real-world environment. Using this context, the computer can help the user to navigate in the real world by presenting a virtual guidepost on the live video of the wearable display [9]. Location sensing can also be used to track every movement of the user and store it as an electronic diary [1], which might be helpful in recalling events, people, and objects seen at a particular location. Visual input from a wearable camera is one of the best sources of information giving location and orientation. In a previous work, Feiner et al. [4] used the Global Positioning System (GPS), magnetometer and inclinometer to find the location and orientation for touring purposes. The use of these sensors, however, was restricted to particular environments. The GPS signals, for example, are blocked by buildings and thus cannot be used indoors. Other previous works, based on computer vision, used artificial markers (fiducials) that were placed in a real-world environment [3][5][7]. Aoki et al. [10] developed a dynamic programming algorithm that uses color histograms to find the trajectory of the user, which works without using fiducials or other sensors. But that method has difficulties in finding the user’s precise location and orientation.

Our approach to finding the user’s location and orientation relative to the real-world environment is based on image-based registration between video frames and a set of images taken beforehand. The overview of the approach is shown in Figure 1. The method uses (1) a set of panoramic images acquired at various points in the environment, (2) annotation s attached to the panoramas, and (3) neighborhood relationships between panoramas as prior knowledge about the environment.

Figure 1: Overview of the approach to finding user’s position and orientation.

In our previous works, we proposed a fast and robust method of image registration using affine transform [11][12]. However, it is generally impossible to align video frames with panoramic images that are
projected onto a cylindrical surface using affine transform. Therefore, the previous method is applicable only if the user’s viewing angle includes the horizontal direction. The method also has a problem in that the difference of lighting condition between video frames and panoramas makes image alignment between them difficult.

To solve the viewing angle problem, we developed an improved method of image alignment between frames and panoramas. The method first transforms the frames to a cylindrical surface using multiple assumptions about the viewing angles of elevation of the frame. Next, the method finds affine parameters between the transformed frames and the panorama for each assumption, and selects the best result that gives the largest cross-correlation value between the aligned frame and the panorama. To solve the lighting condition problem, we use the weighted sum of the absolute difference of brightness and its gradient between the video frames and the panoramas that are robust to changes caused by lighting condition as an evaluation function to be minimized when estimating the affine parameters.

2 Method of image registration

2.1 Overview of the method

The method first reprojects an input video frame to the surface on which the panoramic image is projected. Then, it aligns the re-projected frame and the panoramas that are stored in a database associated with the environmental map. The diagram of the process is shown in Figure 2. The details of each step are described in the following sections.

![Diagram of the process](image.png)

Figure 2: Outline of the proposed method.

2.2 Reprojection onto panorama surface

If the frame is transformed to the surface on which panoramic images are projected, it is possible to align the transformed frame with a panoramic image by using an affine transform.

To reproject the frames to the panorama surface, we need to know the projected point from the center position of the frame to the panorama, even though that point is unknown prior to image registration. The method assumes that the center of the frame is mapped to a point sufficiently close to one of a set of points on the panorama that are chosen so that they then cover the entire panorama at regular intervals. The frame is then re-projected using each point and the affine parameters of image registration between the re-projected frame and the panorama.

![Diagram of re-projection](image2.png)

Figure 3: Re-projection of an input video frame.

As for the panoramic images that are projected onto a cylindrical surface, let \((x_p, y_p)\) be one of points covering the panorama, and \(f_p\) be a focal length of the panorama. The viewing angle \((\alpha, \beta)\) from the focal point is computed using the following equation,

\[
\alpha = \tan^{-1}\left(\frac{y_p}{f_p}\right)
\]

\[
\beta = \frac{x_p}{f_p}.
\]

Using the viewing angle, a point \((x, y)\) on the frame can be reprojected to the point \((x', y')\) on the cylindrical surface by using the following equation. Let \(f_c\) be a focal length of the frame:

\[
x' = f_p \tan^{-1}\left(\frac{x}{y \sin \alpha + f_c \cos \alpha}\right) + f_p \beta
\]

\[
y' = f_p \left(\frac{y \cos \alpha - f_c \sin \alpha}{y \sin \alpha + f_c \cos \alpha}\right).
\]

2.3 Image registration

The method of registration in our previous works [11][12] finds the affine parameters that minimize the sum of the absolute difference of brightness between
images. However, since brightness drastically changes
if the lighting condition does, the method may not
correctly estimate the affine parameters if the frame
is captured at different times. To solve this problem,
we use the sum of the absolute difference of brightness
and its gradient between images to be minimized.

The algorithm for estimating the affine parame-
ters is shown below. Let $I_f(x, y)$ and $I_p(x, y)$ be
the brightness of point $(x, y)$ of frame and panorama,
respectively, and $I'_f(x, y)$ and $I'_p(x, y)$ be the gradient
of the brightness of frame and panorama, respectively.
Since a panoramic image is cyclical in the horizontal
direction, the following equations are satisfied:

$$I_p(x + nw, y) = I_p(x, y)$$

where $n$ is any integer and $w$ is the width of the
panoramic image.

1. Select a set of pixels whose magnitude of gradient
vector $(\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})$ is greater than a threshold.

2. Compute a pseudo-motion vector $(u_p, v_p)$ defined
by the following equations (3)/(4) on the selected
pixels in Step 1. This motion vector is the weighted
sum of that computed from the brightness and
that from the gradient. Let $e$ be a constant and
$w$ be a weight of gradient.

$$u_p = -(1-w) \frac{\partial I}{\partial x} + cw \frac{\partial I'}{\partial x}, \quad (3)$$

$$v_p = -(1-w) \frac{\partial I}{\partial y} - cw \frac{\partial I'}{\partial y}, \quad (4)$$

where

$$\frac{\partial I}{\partial t} = I_p(x, y) - I_f(x, y) \quad (5)$$

$$\frac{\partial I'}{\partial t} = I'_p(x, y) - I'_f(x, y). \quad (6)$$

The motion vector is thought to be robust to brightness
changes caused by lighting condition
since its gradient part is relatively immune to the
lighting influence compared to the case when the
gradient part is neglected (in the case of $w = 0$).

3. Select the pseudo-motion vector computed in
Step 2 that is verified by pixel-wise matching using
the following equation,

$$(1-w)||I_p(x + u_p, y + v_p) - I_f(x, y)|| + cw ||I'_p(x + u_p, y + v_p) - I'_f(x, y)|| < \theta,$$

where $\theta$ is a threshold.

4. Estimate the affine parameters $(a_0, a_1, ..., a_5)$
by statistically fitting the verified motion vectors
$(u_p, v_p)$ to the following equation,

$$x' = a_0x + a_1y + a_2$$

$$y' = a_3x + a_4y + a_5. \quad (8)$$

where $x' = x + u_p, y' = y + v_p$. However, the
motion vectors verified in Step 2 include two types of
outliers: (a) those caused by the objects that are
not present in the panoramic image; (b) those
caused by the wrong pixel-wise matching. We
therefore use the M-estimator for the statistical
fitting to reject those outliers.

5. Compute motion vectors $(u_p, v_p)$ at each pixel by
using the affine parameters estimated in Step 4.
and displace all pixels of $I_f(x, y), I'_f(x, y)$
with the vectors. Then, repeat Steps 2-5 until the
estimated parameters converge or a fixed number of
iterations are completed.

6. Compute the normalized cross-correlations
between $(1-w)I_f(x, y) + cwI'_f(x, y)$ and $(1-w)I_p(x', y') + cwI'_p(x', y')$ for $(x, y) \in I_f$ by using
the affine parameters as a confidence value of the
computed parameters.

By selecting the combination of the panoramic, the
frame, and the image alignment parameters between
them that give the best confidence value, the method can
find the position and orientation of where the frame is taken from.

3 Using inertial sensors

The method of image registration described in Section 2
might fail if the frame contains feature-
less scenes, or if objects that are not present in
the panoramic images account for the majority of the scene.
Furthermore, the computational cost for the
image registration is so large that its processing
throughput is limited. Our current implementation
achieves throughput of around 8-10 frames/s and at
500-800 ms of delay. To overcome such problems and
performance limitations, we used inertial sensors [13].

The inertial sensors, fixed on a camera, can mea-
sure the camera’s rotational angles around three axes.
Let the yaw, pitch, and roll of the rotational angles
between time $t+1$ and $t$ be $(\phi, \theta, \psi)$, the rotation
matrix be $R(\phi, \theta, \psi)$, $r_{ij}$ be an element of the matrix
and $f$ be the focal length of the camera. By setting
$r_{31} = 0, r_{32} = 0$, and $r_{33} = 1$ to approximate by affine
transform, the following transform matrix for image
alignment between the consecutive frames, $I_{t+1}$ and
$I_t$, is obtained as:

$$A_{t+1 \rightarrow t} = \begin{bmatrix}
    r_{11} & r_{12} & r_{13} \hat{f} \\
    r_{12} & r_{22} & r_{23} \hat{f} \\
    0 & 0 & 1
\end{bmatrix}. \quad (9)$$

Combined with the affine transform parameters
computed by the image registration method described
in Section 2, the affine transform matrix $A_{t+1 \rightarrow p}$
for image alignment between input frame $I_{t+1}$ at time $t+1$
and the panoramic image can be computed by using

$$A_{t+1 \rightarrow p} = A_{t \rightarrow p} A_{t+1 \rightarrow t}. \quad (10)$$
where
\[
A_{t \rightarrow p} = \begin{bmatrix}
a_0 & a_1 & a_2 \\
a_3 & a_4 & a_5 \\
0 & 0 & 1
\end{bmatrix}
\]  
(11)

We use the affine transform parameters \(A_{t \rightarrow p}\) if the confidence value of the image registration is below a threshold, or if the process of image alignment is still in progress so that its processing throughput and delay can be improved.

4 Systems and applications

We have constructed a prototype wearable vision system, VizWear [15], as an infrastructure for developing and performing various kinds of visual tasks.

4.1 VizWear system

The system consists of a wearable computer (OS: Windows 98; CPU: Mobile PentiumII-500 MHz; weight: 1.53 kg) equipped with a head-worn display (MicroOptical, Clip-on display), a CCD camera (Toshiba, IK-SM43H), inertial sensors (InterSense, InterTrax\textsuperscript{\textregistered}), a wireless LAN card that complies with 11-Mbps IEEE 802.11b, and a remote PC cluster consisting of 4 PCs (OS: Linux-2.2.14 SMP supported; CPU: Dual PentiumIII) as shown in Figure 4. The wearable PC held by the user captures and compresses input video frames with JPEG encoding and transmits them to the PC cluster via the wireless network. The cluster receives, decodes, and processes the video frames and then sends the result back to the wearable PC. The prototype system also works with other wearable vision applications [14].

![Figure 4: Overview of VizWear.](image)

We implemented the improved method of image registration described in Section 2 and 3 as software. The software of our implementation uses the Parallel Virtual Machine (PVM) library [17] for data distribution and collection among PCs so that it can run independently from architecture and operating system of targeted computers. For high-performance computation, we used parallel computation based on multi-thread programming model: the POSIX thread for UNIX and the Win32 thread for Windows 2000 and Windows 98. The unit of processing is the estimation, from one initial estimate, of the affine parameters of image alignment between a frame and a panorama. In this software, the process of estimation is implemented as a thread code so that it can be speeded up by increasing the number of CPUs and PCs without having to rewrite the code.

4.2 Annotation overlay

Annotation overlay on a live video is an essential feature of augmented reality (AR), since it makes possible a wide range of applications. One of the most promising is an augmented memory application helping users to remind of schedules and “to-do” lists by providing virtual notes overlaid with suitable scenes in the real-world locations such as a workbench desk for “to-do” lists and a refrigerator for shopping lists. Annotation overlay can be also used for personal navigation and touring assistance by showing virtual guideposts and descriptions [8][9].

By placing annotations on the panoramic images as prior knowledge, we can map the position of the annotations from the panorama to the frames by using the image alignment parameters between them, and can thus generate the frames overlaid with the annotations as shown in Figure 5.

![Figure 5: Panorama-based annotation overlay.](image)

5 Experiments

We evaluated our improved method in both off-line and on-line experiments.

5.1 Evaluation of image registration

In the off-line experiment, we captured the video frames (size: 320 \(\times\) 240; 100 frames) and a panoramic image (size: 2560 \(\times\) 480) at a height of 1.5 meter, and we manually gave the correspondence of 10 points from the frames to the panorama. The video frames we used contain objects that were not present in the panoramic image. The panoramic image was created by geometrically transforming an omnidirectional image to a cylindrical surface. The omnidirectional
image is captured by a HyperOmni Vision sensor [16] consisting of hyperboloidal mirror attached on a high definition digital camera (Sony, DSC-F905V; 3.3 million pixels CCD).

We evaluated the method with regard to its image alignment error by using the ground-truth correspondence data. The image alignment error of the improved method and of the previous method (affine transform) are shown in Figure 6. The results show that the improved method reduced the image alignment error by 3–10 pixels. The improvement is especially evident when the viewing angle does not include the horizontal direction (frame number: 50–80). As constants in equations (3) and (4), we empirically set $c = 2.0$ and $w = 0.25$. For a set of seed-points on the panoramic image required for the initial stage of image alignment, we give the seed points at regular 30-pixels intervals, for a total of 768 points.

![Figure 6: Image registration error for previous and improved methods.](image)

**5.2 Evaluation of the robustness to lighting condition**

In the off-line experiment, we compared the results produced by the previous method and those by the improved method in terms of robustness to the lighting conditions. We found that the previous method was prone to produce false estimation under different lighting conditions whereas the improved method was not.

![Figure 7: Examples of two video frames captured in different lighting condition.](image)

Figure 7 shows two video frames captured at the same location under different illumination (in the daytime and nighttime respectively) and Figure 8 show the panoramic image used in the experiment which was captured in the nighttime. The improved method can align the video frames with the panorama, but the previous method cannot align the video frame captured in the daytime with the panorama.

![Figure 8: The panoramic image used in the off-line experiment.](image)

![Figure 9: Distribution of confidence value produced by the previous method (above) and the improved method (below).](image)

The distribution of confidence value for the video frame captured in the daytime from each seed-point produced by the previous and improved method is shown in Figure 9.
The results show that in the distribution of confidence value produced by the improved method, the acute single peak is located at the correct parameters, whereas, in that produced by the previous method, the peak is at false parameters although the sub-peak is at the correct parameters and that causes the false matching.

5.3 Evaluation of position estimation

We conducted on-line experiments to test the accuracy of the user’s position using 45 panoramic images (size: 2560 x 480) captured in the environment shown in Figure 10. Each cross mark on the map represents the position at which a panorama is captured and the distance between cross marks is 50 cm. Portions of the captured panoramic images are shown in Figure 11.

Figure 12 compares the user’s estimated position with the user’s actual position. The result shows that the improved method can acquire the user’s position to an accuracy of two blocks or 100 cm. Because the panoramic images captured at an open space such as the middle of a room give small parallax between them, the frames taken from them can be aligned to any of the panoramic images with high confidence values, thus complicating precise positioning. However, as long as the same annotations are placed to these panoramic images, the output frames overlaid with correct annotations can be generated. Examples of output video frames overlaid annotations and the map are shown in Figures 13 and 14. Although the estimation result of the user’s position fluctuated, the annotations were stably overlaid. We also found that the improved method could robustly find the correct image alignment for the images captured at different times, even in the daytime and nighttime.

The computational cost required for image registration between a single frame and a set of panoramic images was 30 ms per seed points or total 400–800 ms. By parallelizing the image registration among the PC cluster, the total cost was reduced to 80–160 ms or 6–12 frames/s and the delay was 600–800 ms. With inertial sensors, the total throughput and delay were improved to 10–15 frames/s and 100–150 ms, and therefore real-time processing can be achieved. Our method at the initial state requires thorough frame-to-panorama image registration and thus its computational cost is proportional to the number of panoramic images. In the experiments, the user is assumed to be near the entrance in Figure 10 and thus the number of panoramic images is reduced to 9. The computational cost required for the initial thorough search is 5–10 seconds.

6 Conclusion

We have developed an improved method of image registration to find the location and orientation of the user of a wearable computer. The method reprojects the input video frames to the surface on which panoramic images are projected to enable image registration with an affine transform. Experimental results show that the accuracy of image registration is improved by 10–15% without greatly increasing computational cost. We also demonstrated the application of annotation overlay and personal navigation for wearable computers in on-line experiments with the prototype wearable vision system, VisWear, that is composed of off-the-shelf hardware components. In the future work, we need to improve time-consuming
process at the initial state of the method using different clues such as color histograms.

Acknowledgements

This work was conducted as a part of the Real World Computing (RWC) Program.

References


