# Camera Tracking in the Process of Designing Selective Auto-Darkening Filter

**Sven Buchholz, Arsalan Malik, Bernd Hillers and Axel Gräser**

**Friedrich-Wilhelm-Bessel-Institut (FWBI), 28063 Bremen**
**{sbuchholz, amalik, bhillers, ag}@iat.uni-bremen.de**

**Keywords:** *augmented reality, manual arc welding, homography-based tracking, head-mounted camera tracking, welding helmet, marker-less tracking, SIFT features.*

**Abstract**

*Welding is a very important industrial process. Eye protection of welders is nowadays achieved by helmets in which a so-called auto-darkening filter (ADF) is integrated. Physically, an ADF is a one block LCD that can switch to a pre-selected shade level at the moment a welding arc is struck. This global darkening, however, is not optimal. A selective ADF (SADF) that only darkens the area where the welding arc appears, would be superior to an ADF. We demonstrate how such a SADF can be achieved, in principle, with a single head-mounted camera (HMC) as sensory input. The key idea is to do the camera calibration in an arbitrary but fixed plane that initially contains the welding arc, and, then to exploit the fact that the mapping between the HMC and the SADF is always a homography. These two things together allow us to successfully perform inside-out tracking of the HMC using only image information. Our proposed tracking algorithm is marker-less and uses SIFT for the computation of image features. Many aspects of the algorithm are detailed and matching and tracking results are presented. A simple outlier detection scheme that reduces the number of mismatches is also introduced.*

## 1 Introduction

Welding is a very important industrial process. Though robot welding is becoming more commonplace in the industry, there are still more than 45 million welders worldwide. Todays welding helmets provide very effective eye protection to the welder. State-of-the-art are so-called auto-darkening filter (ADF). Physically, an ADF is a one block LCD that can switch to a pre-selected shade level at the moment a welding arc is struck. On the other hand, the helmet restricts the welder in his or her control of the welding process, e.g. by limiting the filed of view.

IntARWeld (Intelligent Augmented Reality Welding Helmet) is a project funded by the European Commission and carried out together by FWBI and Sperian. The project aims for overcoming the limitations of current welding helmets. Its goal is to provide high-tech helmets with very new, intelligent eye protection and information systems for welders.

In this paper a system that advances current ADFs is proposed. An ADF always delivers a global homogeneous darkening. This way, not only the welding arc gets shaded but also the whole environment. A selective ADF (SADF) that only darkens the area where the welding arc appears, would be superior to an ADF. On the technical site, this

requires the replacement of the one block LCD with a graphical LCD (GLCD) with high pixel resolution, very fast response time and high contrast between clear and dark states. Designing a SADF has also very interesting perceptional aspects.

Here, however, we show how to solve the geometrical aspects of the problem. We demonstrate how a SADF can be achieved, in principle, with a single head-mounted camera (HMC) as sensory input. The key idea is to do the camera calibration in an arbitrary but fixed plane that initially contains the welding arc, and, then to exploit the fact that the mapping between the HMC and the SADF is always a homography. These two things together allow us to successfully perform inside-out tracking of the HMC using only image information.

The next section gives an overview of our system and its calibration. In the third section the homography-based inside-out tracking algorithm for the HMC is derived. Point correspondences are obtained from applying SIFT as detailed in section four. Matching and tracking results are presented in section five and the paper then concludes with a summary and an outlook.

## 2 Related Work

A wearable Augmented Reality system TEREBES [1] for the observation of welding process has been developed [1]. The welding scene is captured by two high dynamic range cameras mounted inside helmet. The welding scene is enhanced to improve the contrast, and then augmented with a user interface. The resultant scene is rendered in a video see-through head-mounted display. The system is able to solve the visibility problems. However, the disadvantage is that both the computational complexity and cost of the system are very high. Moreover, the ergonomics evaluation of TEREBES suggests that users have difficulties working with video see-through displays due to issues such as quality, displacement of cameras from eye position, field of view and most importantly time lag [2].

The calibration of optical see-through is challenging and its accuracy is dependent on skills of its user. Many researchers have used different methods for calibration depending on the application and required accuracy of the system [3] [4] [5] [6]. Tuceryan et. al [7] presented an alignment method based on single point in the world. The single point is tracked using 6 degrees of freedom (DoF) magnetic tracker, while its corresponding point on HMD is reported by the human observer. A technique based on camera calibration is reported by Gilson et. al [8]. They use a camera instead of human observer for a fully automated calibration procedure. However, they also use markers to track the position of head. The calibration technique presented in this paper does not require any marker tracking.

## 3 Calibration

As HMC a low-cost camera from the German vendor VRmagic has been used. This camera, a VRmC-12, has a dynamic range of 80 dB and can deliver color images of $752 \times 480$ pixels at a frame rate of about 28 fps. The other component of the assembled prototype helmet is a see-trough graphical LCD (GLCD). This GLCD has a refresh rate

[1] Tragbares Erweitertes Realitäts-System zur Beobachtung von Schweißprozessen

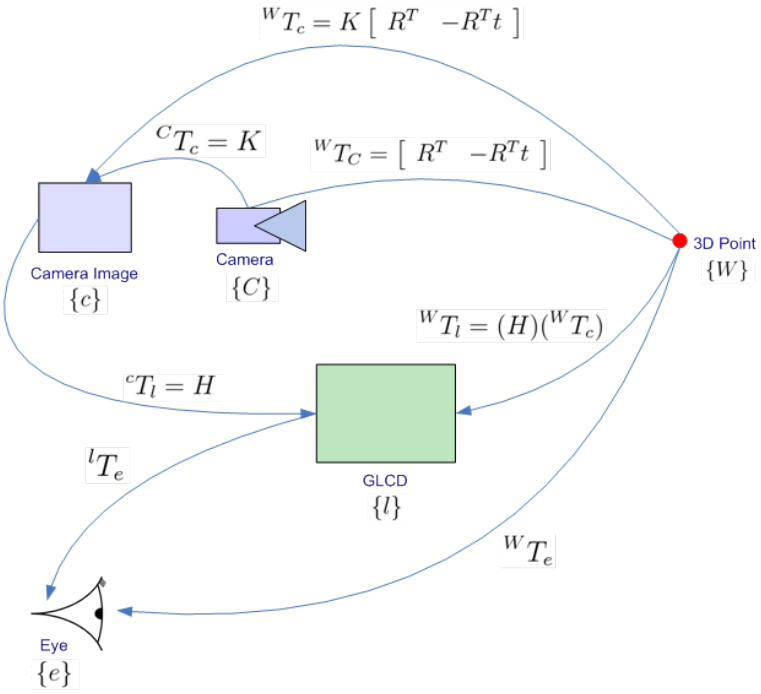


**Figure 1:** The transformations that describe the SADF problem geometrically.

of only 6 fps and very low contrast. Its resolution is only $147 \times 132$ pixels. With these values, the prototype can not be used by a human for actual welding.

For the realization of the SADF, we need to know how a world point is perceived on the GLCD. Our only available tool for solving this problem is the camera image of the HMC. Geometrically, the situation is described by the set of transformations as given in figure 1.

Using a camera as a sensor requires its calibration first. For a camera, assuming the pinhole-model (see e.g. [9]), this comes down to the computation of the so-called projection matrix $P$, which describes the transformation from world coordinates to image coordinates (pixels). The matrix $P$ can be decomposed into two transformations. The first is an Euclidean motion formed by a rotation matrix $R$ and a translation vector $t$. These are the so-called external parameters of the camera which are also often referred to as it pose (position and orientation w.r.t. the world coordinate system). The second transformation is described by the so-called internal calibration matrix $K$. If we move the camera, $K$ remains constant and only the external parameters of the camera have to be recalculated.

Additionally to camera calibration, the transformation from the camera image to the GLCD has to be computed as well. This is a mapping, say $H$, from one plane to another one, i.e. a homography. Technically, a homography is a invertible square matrix of size three. Obviously, $H$ is another non-varying component of our system.

Recovering the camera pose from one single image is not possible - due to the lost of depth during the process of image forming. On the other hand, no lost can occur if there is no depth in the scene. This means that for a planar scene the camera projection matrix reduces to a homography (the one from the world plane to the image plane). Treating a non-planar scene as being planar results in an approximation error (see e.g. [10]) and

the concept of the relative affine structure of a scene [11]. Now the welding arc can be modeled as a set of points lying on a particular world plane. We coin this plane the welding arc plane $\{W\}$. Only the points of this plane have to be mapped without error to the GLCD for selective auto-darkening.

All of the above suggests to do the calibration of the system as follows. Start with determining the internal parameters $K$ by any standard calibration method. Then collect point correspondences by the following procedure, which is an adaption of the method proposed in [7]. Move a LED (simulating the welding arc) in (approximately) a plane at fixed depth from the user's eye, display a dark square at random position of the GLCD and ask the user to cover the LED with it by moving his or her head. This gives one correspondence between the position of the LED in camera image and its position on the GLCD. The position in the camera image can be obtained by means of simple intensity-based tresholding.

Given a set of such point correspondences $p_i \Leftrightarrow p'_i$ in homogeneous coordinates a solution for $H$ can be derived from

$$p'_i \times Hp_i = \mathbf{0} \tag{1}$$

where $\mathbf{0}$ stands for the null vector of appropriate dimension. Each point pair $(x, y, 1) \Leftrightarrow (x', y', 1)$ gives two linear equations

$$x'(h_{31}x + h_{32}y + h_{33}) = (h_{11}x + h_{12}y + h_{13}) \tag{2}$$

and

$$y'(h_{31}x + h_{32}y + h_{33}) = (h_{21}x + h_{22}y + h_{23}) \tag{3}$$

with $h_{ij}$ being the entry of $H$ at position $(i, j)$. This leads to a system of the form

$$A\mathbf{h} = \mathbf{0} \tag{4}$$

where $\mathbf{h}$ is just the components $h_{ij}$ stacked into a vector.

A homography has eight degrees of freedom. This is because it describes a projective relation and is therefore only defined up to scale. Hence an exact solution is obtained from four correspondences. More than four point pairs give an overdetermined set of equations. A solution for (4) can be obtained from applying the Direct Linear Transform algorithm [9]. It is advisably to always use more than 4 point pairs since all correspondences are noisy measurements. Usually, with 12 correspondences we always managed to get a calibration error (calculated by back-projection) of less than one pixel on the GLCD.

# 4 Tracking Algorithm

The system is calibrated for a virtual plane at fixed distance from the welder's eye. As long as the welding arc (simulated by the LED) does not leave this plane, the darkening will take place at the right positions on the GLCD. Otherwise, a registration error will result. For quantifying this error the following setup has been developed. A second camera —coined the virtual observer camera (VOC)— captures the GLCD from inside of a phantom's head, where it is positioned at the place of the right (or the left) eye. Describing the calibration of the VOC, i.e. the computation of the non-varying mapping

from the VOC image to the GLCD, is omitted here. After calibration, the position of the LED on the GLCD under this mapping serves us as ground truth.

Figure 2 shows the registration error for several runs of measurement. The error is already about 6 pixels when the system is moved 10 cm away from this plane. It can be seen that the registration error grows linearly with the distance from the calibration plane. The registration error grows linearly with the distance from the calibration plane, which can also be verified by a simple calculation. Registration errors due to movements can be compensated by means of tracking.

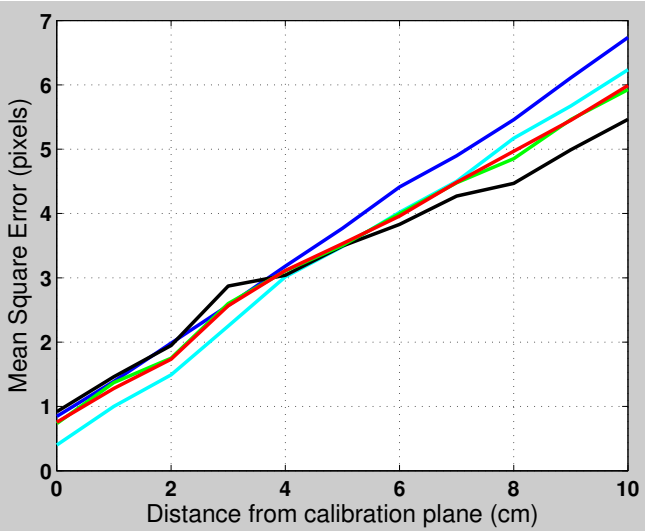


**Figure 2:** The registration error as a function of the distance to the calibration plane.

Tracking can be done in two ways: either by using artificial markers, or by relying on natural image features only. All the benefits of markers (easy to detect objects of known size become available for tracking) come at one price. One simply has to be able to engineer the environment in such a way first. Augmenting a generic welding scene with large enough markers that are sufficiently visible but do not distract the welder is not possible. Hence we headed for marker-less tracking for the IntARWeld project.

An object that is likely to be visible in any HMC image is the welding pistol. The contour of the burner can be viewed as being approximately formed of two parallel line segments. Nevertheless, detection of the pistol using standard methods such as the Hough transform very often failed. Finding suitable image features for tracking is detailed in the next section. In the reminder of this section we introduce our new tracking algorithm.

Recall that registration errors result from leaving the calibration plane with the welding arc. More precisely, the plane containing the welding arc does no longer coincide with the latter. Nothing more than this new position of the welding arc plane (WAP) is needed for the realization of the SADF. Thus it is sufficient to track the movement of the WAP. This movement is a homography, which can be determined in the image domain by using point correspondences from image pairs in exactly the same way as shown in the previous section. In fact, we designed the calibration of the system in such a way that it delivers an initial solution for tracking the WAP.

Plane tracking in general, or, equivalently homography-based tracking, is well understood (see e.g. [12, 13] and references therein). In the following we apply the standard

algorithm to our setup.

Specifically, we set the calibration plane to be the reference world plane at $z = 0$. Calibration of our system yields the matrix $H_w^0$, which is the planar homography that maps maps points on the world plane to the image plane of the calibrated camera. The latter acts as our reference frame 0. Again, note that $H_w^0$ is the camera projection matrix for points on the plane. The so-called inter-image homography $H_i^{i+1}$ for two consecutive frames $(i, i+1)$ can always be computed from point correspondences. Therefore we obtain the following chain

$$H_i^{i+1} = H_{i-1}^i H_{i-2}^{i-1} \dots H_0^1 \tag{5}$$

and finally, using the value of $H_w^0$

$$H_i^w = H_0^i H_w^0 \,. \tag{6}$$

As we know, every projection matrix can be decomposed into its internal parameters ($K$) and external parameters (pose). Performing this decomposition for our case yields

$$H_i^w = K \, [r_1^i \, r_2^i \, t] \tag{7}$$

with $r_j^i$ ($j \in \{1, 2\}$) being the $j$-th component of the rotation matrix and $t$ being the translation vector. $K$ is known from calibration and hence we can extract $r_1^i$ and $r_2^i$ from the first two colums of $K^{-1} H_w^i$. The third component of the rotation matrix is given by the cross product $r_1^i \times r_2^i$. Thus we can compute the pose using only inter-image homographies if we know both $K$ and $H_w^0$. Homography decomposition in all its details is treated in [14]. To complete one update step of the algorithm from one frame to another, the fixed and predetermined homography from the image plane to the GLCD (see section two again) is finally performed.

# 5 Feature Detection using SIFT

In the absence of markers, the detection of robust image features and their reliable matching becomes a very challenging problem. Moreover, any movement of the welder's head induces a potentially large change in the viewpoint between two consecutive images taken by the HMC. Hence matching must be invariant to image transformations and illumination changes. Note that in a welding scene huge illumination changes do occur.

A well–established algorithm that fulfills all these requirements is the Scale Invariant Feature Transform (SIFT), which has been introduced by David Lowe in [15]. SIFT first detects and localizes feature points in different scale space images. Then an orientation is additionally assigned to every localized feature point. The combination of a feature point and a corresponding orientation is a so–called keypoint. For each computed keypoint a descriptor is finally assembled using orientation histograms. In [16] the authors carried out a comparative study on the performance of many different local image descriptors. According to this study, the SIFT approach performs best. Hence it was chosen for implementation. Specifically, a freely available C-library made by Rob Hess has been migrated to C++.

SIFT constructs a discretized Gaussian scale space. There are two discretization parameters. The first parameter is the number of so-called octaves $O$. Each octave is further

subdivided in $S$ sub–levels. The scale corresponding to octave $o$ and sub–level $s$ is given by the formula

$$\sigma_{o,s} = \sigma_0 \, 2^{(o+s)/S} \,, \tag{8}$$

where $\sigma_0$ is the smoothing for the base level. For each pair $(o, s)$ the current image is convoluted by an isotropic Gaussian kernel of variance $(\sigma_{o,s})^2$. Additionally, image size is reduced by a half at each successive octave. Consequently, the image size at the base level is the dominant factor of the time complexity of SIFT. Hence computation time grows linear with the number of pixels of the base level image.

Note that it is best practice to determine the number of octaves $O$ in dependence of the image size as follows:

$$O = \log(\,\min\,(\mathit{image\ width},\, \mathit{image\ height}))\,/\,\log(2) - 2\,. \tag{9}$$

According to the literature, the smallest reasonable number of sub-levels is three. With three sub-levels and five octaves feature detection takes about 1.2 seconds on a Core 2 Duo P8700 for an image of $752 \times 480$ pixels, which is the native resolution of the VRmC-12 camera.

Of course, the number of detected features varies with scene contents and scene contrast. The fact that the welding torch, the welder's hand and the workpiece will be (partly) visible in the scene should guarantee for sufficient inhomogeneous image parts. Feature detection has been evaluated for various different sequences and varying contrast. On average, 80-100 features are detected. For a very low contrast image this number may drop as low as ten features. An example of feature detection for a low contrast frame is shown in the left panel of figure 3. The right panel of this figure shows a frame yielding about 200 features.

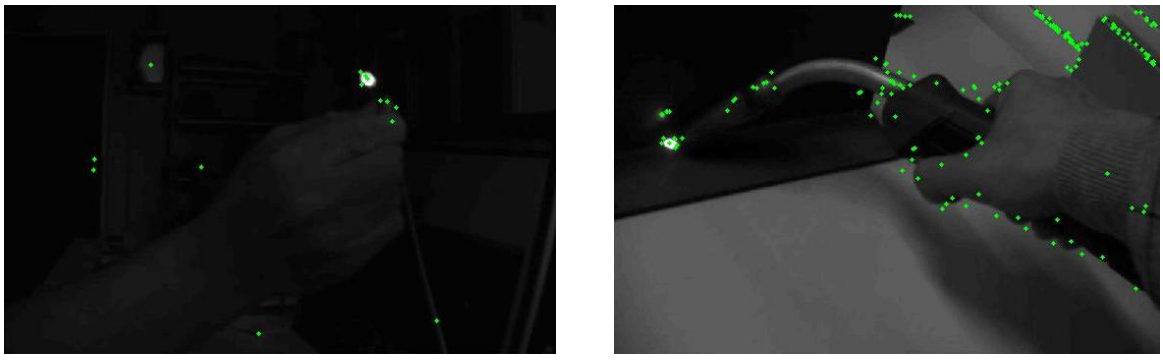

**Figure 3:** SIFT features for two frames with different contrast.

The only way to reduce detection computation time is to reduce the image size. We therefore studied the impact of the image size on the number of detected features. Results for one typical welding scene (normal contrast setting) are reported in figure 4. On average, reducing the image size from $752 \times 480$ pixels to $564 \times 360$ pixels results in a lost of 10%-15% of the features. This a very acceptable rate. For a size of $376 \times 240$ the rate is in the range of 20%-35%. This might still be acceptable having in mind that computation is speed up by a factor of four. For a resolution of $188 \times 120$ pixels the number of detected features is usually too small for further processing, i.e. reliable matching.

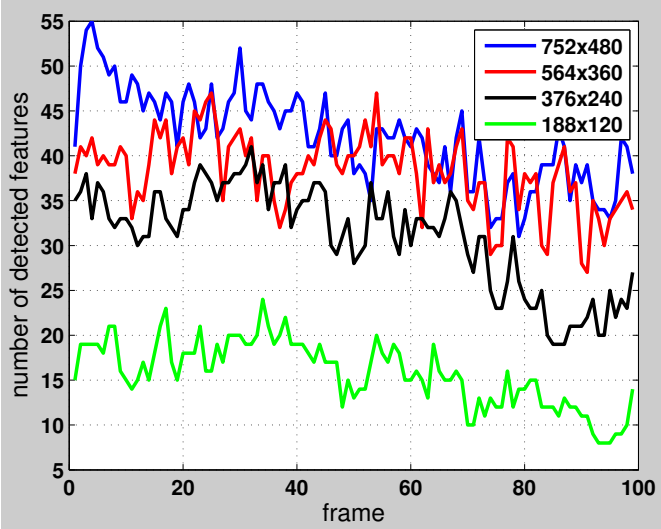


**Figure 4:** Impact of the image size on the number of detected features.

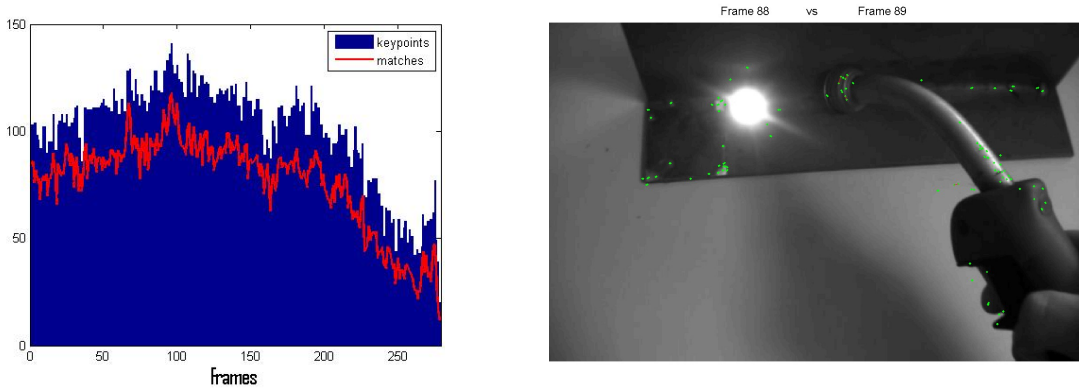


**Figure 5:** Matching consecutive frames of a typical welding sequence. Left panel shows the result of always matching against the previous frame of the sequence. Right panel shows the 88th frame and the 89th frame of the sequence overlaid in one single image. Matched features are connected by green lines.

# 6 Matching and Tracking Results

For a few hundred features, feature matching needs only little time compared to feature detection. Therefore it is affordable to perform the matching between two frames in "both directions". This means a match is only accepted if it is found by both matching the first image against the second one, and vice versa.

Theoretically, always matching the current frame against the previous one is optimal. If this is done for a typical welding scene, the number of matchings remains constantly high over time and will follow any trend that is present in the variation of the number of detected features. This can be observed from the left panel of figure 5. Practically, however, this sampling with the full time resolution of the camera may be unnecessary. The right panel of figure 5 shows two consecutive frames of the sequence overlaid as a single image (with the detected features). There are hardly any differences notable.

It is very likely that for manual welding the relevant movements are mostly slow and

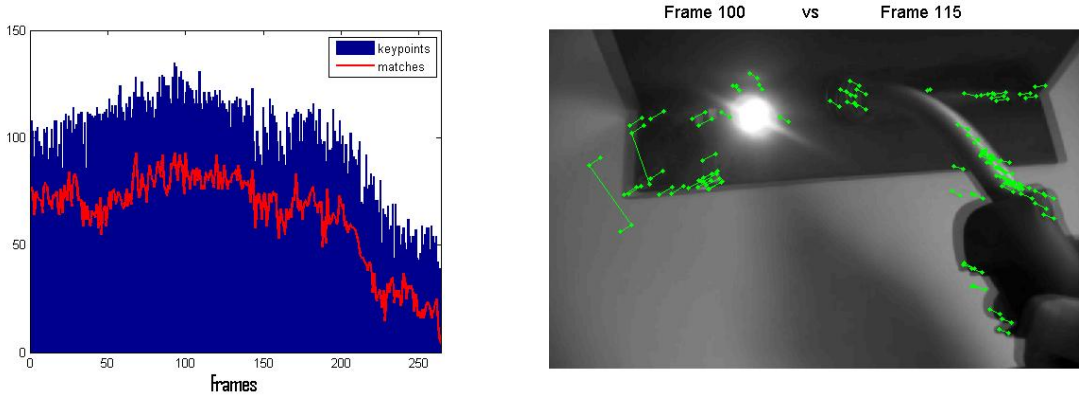


**Figure 6:** Results for matching with an offset of 15 frames. Left panel depicts the number of detected features and the corresponding number of matches for the whole sequence. Right panel shows the overlay of two frames. Matched features are indicated by green points and are connected by a green line.

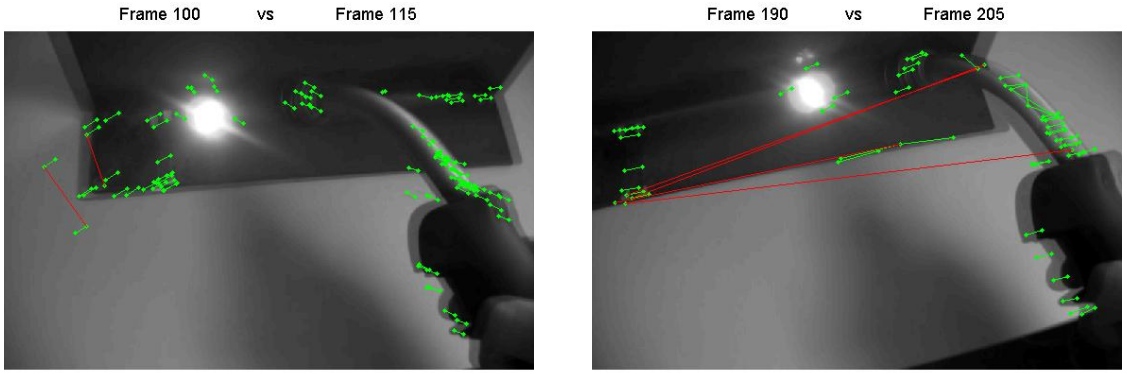


**Figure 7:** Outlier detection. Rejected matches are connected by a red line.

smoothly. Hence frames may be skipped for tracking. In order to get an estimate of how many frames can be left out MATLAB routines for automated analysis of image sequences have been developed. It seems possible to match frames with a time offset up to 15 frames without loosing accuracy. An illustration of the effect of applying this scheme to a particular sequence is given in figure 6. As can be seen from the left panel of this figure no (independent) breakdown of the number of matches did occur.

Of course, more important than the number of found matches is the number of those matches, which are actually correct ones. For robust matching a simple outlier detection method has been implemented. The method first computes the length of all matching lines in the overlaid image. Then the sample mean $\mu$ and the sample standard deviation $\sigma$ of this data are computed. All matches for which the computed length of the matching line does not lie within $[\mu - 2\sigma, \mu + 2\sigma]$ are rejected. For illustration purposes, results obtained with this method are presented in figure 7. Since the method works fast it should be applied twice. This way, outliers that may have been "overlooked" in the first run (due to the presence of dominating outliers) may be detected in a second run.

All remaining point matches are used to compute the inter-image homography. The computation involves the application of a RANSAC scheme. The support of the homogra-

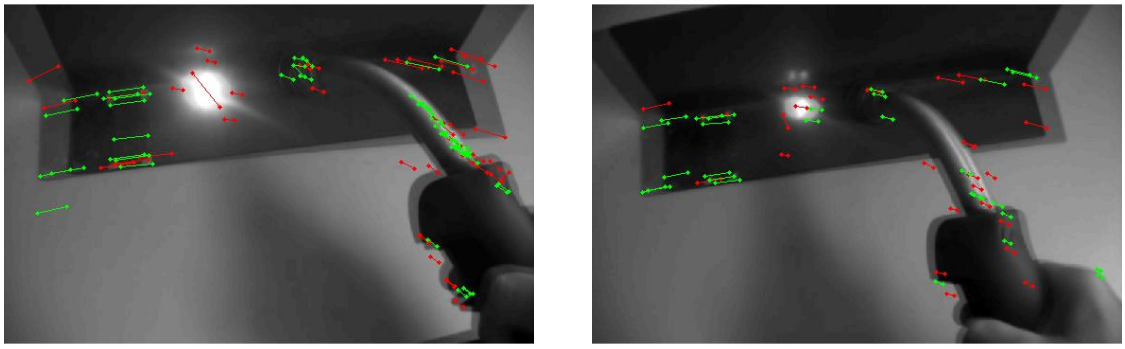

**Figure 8:** Homography support. Matches that support the estimated homography are shown in green while the ones that do not support it are shown in red.

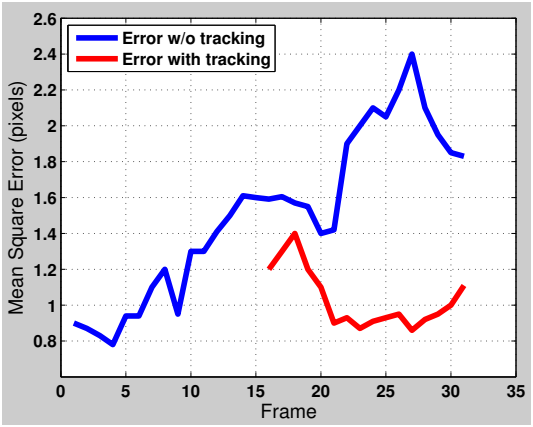


**Figure 9:** An example of the performance of the tracking algorithm. The blue curve shows the error without updating the homography trough tracking. The purple curve shows how the error is reduced after performing an update resulting from matching the 15th frame of the sequence against the 1st frame.

phies computed this way has been found to be in the range of 36% to 74%. These numbers are quite reasonable. Even after outlier rejection, mismatches will remain undetected since our simple method is not perfect. Such mismatches will therefore also contribute to the estimation of the homography. In fact, they may well be the dominant factor since it looks more attractive to RANSAC (in terms of minimizing the error criteria) to find a homography that supports the mismatches. Illustrative examples of homography support are given in Figure 8. For consecutive frames the support rate should is nearly 100%. This indicates that this rate can be used for calculation if a time offset is too large. An important point to notice is that usually all image areas contribute to the homography.

Several experiments have been performed in order to evaluate the performance of the tracking algorithm. The results look promising but are still preliminary. This is due to the fact that performance is judged by applying other image processing algorithms, i.e. without having a ground truth. The registration error is determined on the GLCD. Therefore the camera images have to be downsized to the much lower resolution of the display. This is of course harmful to a precise evaluation. Figure 9 gives one example of the performance of the algorithm.

# 7 Conclusion

We showed how a welding helmet with a selective auto-darkening filter can be realized, in principle, with only a single head-mounted camera. Specifically, we developed the therefore needed mathematical tools in terms of calibration and tracking. Our whole approach centers around the fact that only the world points constituting the welding arc have to be mapped without error to the GLCD in order to achieve the proper selective auto-darkening. This directly led us to homography-based methods, i.e. a solution that is solely formulated in 2D.

Since augmenting a welding scene with artificial markers is not practically possible we implemented a marker-less tracking. Image features have been generated using SIFT because of its known high reliability w.r.t. image transformations and illumination changes. First result are promising but have to be put in the context of the available hardware. The GLCD of the prototype has only a low resolution. Hence registration errors are measured on a rather coarse scale. Moreover, they could only be measured by using another camera as ground truth, which can not be an exact way by definition.

# References


[1] Aiteanu D. Hillers B. and A. Gräser. Augmented-reality helmet for the manual welding process. *Virtual and Augmented Reality Application in Manufacturing*, pages 361–381, 2004.

[2] M. Park, L. Schmidt, C. Schlick, and H. Luczak. Design and evaluation of an augmented reality welding helmet: Research articles. *Hum. Factor. Ergon. Manuf.*, 17(4):317–330, 2007.

[3] Gang Luo, Noa Rensing, Evan Weststrate, Eli Peli, and Member Spie. Registration of an on-axis see-through head mounted display and camera system. *Optical Engineering*, 2005.

[4] Yasuyoshi Yokokohji, Yoshihiko Sugawara, and Tsuneo Yoshikawa. Accurate image overlay on video see-through hmds using vision and accelerometers. *Virtual Reality Conference, IEEE*, 0:247, 2000.

[5] Hirokazu Kato and Mark Billinghurst. Marker tracking and hmd calibration for a video-based augmented reality conferencing system. In *IWAR '99: Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, page 85, Washington, DC, USA, 1999. IEEE Computer Society.

[6] Michael Figl, Christopher Ede, Wolfgang Birkfellner, Johann Hummel, R. Seemann, and Helmar Bergmann. Calibration of an optical see through head mounted display with variable zoom and focus for applications in computer assisted interventions. In *In Medical Imaging 2003: Visualization, Image-Guided Procedures, and Display, number 5029 in Proceedings of the SPIE*, pages 618–623, 2003.

[7] M. Tuceryan, Y. Genc, and N. Navab. Single-point active alignment method (spaam) for optical see-through hmd calibration for augmented reality. *Presence: Teleoper. Virtual Environ.*, 11(3):259–276, 2002.

[8] Fitzgibbon A.W. Gilson S.J. and Glennerster A. Spatial calibration of an optical see-through head mounted display. *Journal of Neuroscience Methods*, 173:140–146, 2008.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2nd edition, 2003.

[10] M. Irani and P. Anandan. A unified approach to moving object detection. *PAMI*, 29(6):577–589, 1998.

[11] A. Shashua and N. Navab. Relatuve affine structure: Canonical model for 3d from 2d geometry and applications. *PAMI*, 18(9):873–883, 1996.

[12] G. Simon, A. Fitzgibbon, and A. Zisserman. Markerless tracking using planar structures in the scene. In *International Symposium on Augmented Reality, ISAR 2000*, pages 120–128, 2000.

[13] M. Lourakis and A. Argyros. Chaining planar homographies for fast and reliable 3D plane tracking. In *18th International Conference on Pattern Recognition, ICPR 2006*, volume 1, pages 582–586, 2006.

[14] O. Faugeras and F. Lustman. Motion and structure from motion in a piecewise planar environment. Technical Report Rapports de Recherche 865, INRIA, France, 1988.

[15] D.G. Lowe. Distinctive image features from scale-invarinat keypoints. *IJCV*, 60(2):91–110, 2004.

[16] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.