DETECTING HANDS IN VIDEO IMAGES USING SCALE INVARIANT LOCAL DESCRIPTORS

Jan Richarz, Thomas Plötz and Gernot A. Fink
Robotics Research Institute, Intelligent Systems Group
University of Dortmund
Otto-Hahn-Str. 8
44221 Dortmund, Germany
email: {jan.richarz, thomas.ploetz, gernot.fink}@udo.edu

ABSTRACT
In this paper, we describe our approach on hand detection in cluttered images using scale invariant features. We claim that, while modelling hands as a whole is bound to fail because of their strongly articulated nature, treating them as a collection of weakly connected characteristic regions seems promising. Different approaches to finding and robustly modelling such regions - or local object descriptors - invariantly to scale and orientation of the object in question have been proposed. As an example, we demonstrate our approach using the well-known scale-invariant feature transform (SIFT), combined with a region-based postprocessing to eliminate false positives. We present detailed results on a large set of images from a realistic interaction scenario with a smart room.

KEY WORDS
hand detection, human-computer interaction, scale-invariant features

1 Introduction

Our research is focused on the fields of pattern recognition and multi-modal Human-Machine Interaction (HMI). As an integration scenario, we are developing an "Intelligent House" – the FINCA (Flexible Intelligent Environment with Computational Augmentation) [1] – which is equipped with different types of sensors, including active pan-tilt zoom cameras and microphones. Furthermore, the integration of programmable actors (e.g. for lighting and sun-blinds) in the system architecture enables us to control different features of the house via software. Control of these features by the users should take place naturally and intuitively, i.e. using speech and gestures. So, we are particularly interested in gesture-based interaction.

Since gestures are mostly defined by hand/arm poses and motions, a fundamental prerequisite to gesture recognition is the robust detection of hands in images. Hands are strongly articulated objects, and therefore, pure model- or appearance-based approaches, taking into account the object as a whole, are not well-suited for this task. Instead, we aim to describe hands as a collection of characteristic regions – or parts – making the detection robust against articulations and occlusions. In this paper, we describe our approach using a scale-invariant salient feature detector.

The remainder of this paper is organized as follows: First, we present related work and outline the considerations that lead us to our approach, which we describe in Section 3. Section 4 gives an overview of the data and experimental setup we used for deriving the evaluation results presented in Section 5. We conclude with a summary and an outlook on future work.

2 Related Work

Vision-based gesture recognition has attracted a lot of attention in recent years. Therefore, a great number of different approaches to hand and limb detection exist. A straightforward and simple approach that is often utilized (e.g. [2],[3],[4]) is to look for skin-colored regions in the image. Although very popular, this has some drawbacks. First, skin color detection is very sensitive to lighting conditions. While practicable and efficient methods exist for skin color detection under controlled (and known) illumination, the problem of learning a flexible skin model and adapting it over time is challenging. Secondly, obviously this only works if we assume that no other skin-like objects are present in the scene (or that these objects can easily be identified and rejected). So, although skin-color detection is a feasible and fast approach given strictly controlled working environments, it is difficult to employ it robustly on realistic scenes.

Generally, approaches that model an object by its shape, boundaries or general appearance (e.g. the well-known appearance-based object detector of Viola and Jones [5] or Cootes' and Taylor's Active Appearance Models [6]) do not seem well suited for our task. Hands are complex objects having many degrees of freedom in rotation and deformation, thus showing a large variety of shapes. Describing the appearance of a hand with a single, flexible model would result in a model too general for reliable detection. On the other hand, using distinct models for different hand postures would result in a huge model database. We could reduce the complexity of the problem by only looking for a reduced set of predefined hand postures – which is often done in the field of gestural control (see e.g. [3],[7]) – but

this would seriously limit the "naturalness" of our interface.

Recently, there has been increased interest in approaches working with local invariant features. The idea behind is that, if it is possible to identify characteristic points or regions on objects, an object can be represented as assembly of these regions, i.e. rather than modelling the object as a whole, one models it as a collection of characteristic parts (and maybe their spatial relationship). This has the advantage that partial occlusions of an object can be handled easily, as well as considerable deformations or changes in viewpoint. As long as a sufficient number of characteristic regions can be identified, the object may still be found. Therefore, these approaches seem rather promising for the task of hand detection.

In order to reliably find such characteristic regions in realistic scenes, we need keypoint detectors and local region descriptors (i.e local features) that are invariant to position, scale and rotation of the object they describe. Furthermore, they should be (at least to some degree) invariant against affine distortions and illumination changes. Several approaches for keypoint detection and local feature extraction have been proposed, the best-known being probably the Harris corner detector [8] - which, however, is not scale invariant. More recent approaches which are invariant to scale and rotation include the salient regions detector of Kadir and Brady [9], the Speeded Up Robust Features (SURF) proposed by Bay *et al.* [10], and Lowe's Scale Invariant Feature Transform (SIFT) [11].

3 Our approach

We are interested in detecting hands in realistic scenes using scale-invariant local region features. These features - or descriptors - are extracted automatically from the input images using a salient key point detection method, and then matched against a large database of example descriptors using enhanced nearest-neighbor (NN) matching. Since we expect this approach to yield a large number of false positives (which is verified by our experiments), we develop a candidate filtering algorithm based on local neighborhoods that discards a considerable number of false matches.

Feature extraction: Among several possible methods for the extraction of region descriptors, in this paper we concentrate on the SIFT approach for several reasons. First, it is well-known, and its potential has been shown by its successful application in different fields (e.g. rigid object recognition [11], camera calibration and scene reconstruction [12] and localization [13]). Then, it is a staged approach that renders the possibility to modify - or exchange - different stages according to the task at hand. It also is efficient and has potential for considerable speedup.

The first stage of SIFT is the detection of salient keypoints. For this purpose, a Gaussian Scale Space of the input image is constructed. Keypoints are detected as local extrema of Difference-of-Gaussian (DoG) filters in this scale space.

In the next stage, the keypoints are sub-pixel interpolated. Points showing low contrast or being located on edges are discarded, for they are not stable.

Then, each keypoint is assigned a scale (depending on the level of the scale space pyramid it was detected in) and an orientation (according to the principal orientations of gradients in a region around the keypoint).

The final step is the calculation of the local image descriptor, which is a smoothed histogram of gradient orientations and magnitudes of the local image region. In Lowe's original implementation, the descriptor is a 128-dimensional feature vector (see [11] for details). We use a MATLAB/C implementation of the SIFT algorithm provided by Andrea Vedaldi [14].

Matching: Given an image of the object, we now have a number of SIFT features describing it. Since we want to detect hands in different configurations, we build a database containing descriptors extracted from many different hand images (see section 4 for details on the data). The task is now to search for key points in the input images and – based on the saved descriptors in the database – decide whether they are located on hands. Following Lowe's proposal for object recognition, we implement an enhanced NN matching algorithm: Given a feature point that is to be classified, we search the nearest neighbor and the second nearest neighbor that is known to come from a different object. In our application, we have a two-class problem (hand vs. background), so this means we also build a database of SIFT descriptors from background images. The classification score s_{class} is the ratio of the two respective distances d_{pos} and d_{neg} to the nearest neighbors in both databases. By thresholding this score, we classify the descriptors as hand or background.

Candidate filtering: Unfortunately, this simple matching scheme yields a large number of false positives. To reduce this number, we make the assumption that, typically, there will be multiple key points found on hands of which several will be classified as positives, whereas most false positives will be surrounded by negatives. So, to further evaluate each candidate key point labeled as positive by the NN matching, we implement a hysteresis-like rejection criterion based on local neighborhoods.

The first step is to determine a list of keypoints spatially connected with the candidate point in question. We then evaluate points based on the total number of positive candidates in their respective list (threshold $n_{pos,min}$), the ratio between positives and negatives (threshold f_{min}), and the number of positives m they are connected to. The detailed algorithm is shown in Fig. 1. We implemented three different approaches for determination of the keypoint list that is used as input for the algorithm: Circular regions of fixed size, the center being the candidate that should be evaluated, circular regions with sizes proportional to the scale σ of the candidate, and taking the n spatially nearest neighbors. Results for all three approaches will be presented in the next section.

Figure 2. Some typical examples of camera images used for the presented approach.

for all keypoints $k_1 ... k_n$ labeled as positives **do**
  $list$ = getKeypointList(k_i)
  count positives n_{pos} and negatives n_{neg} in $list$
  if $(n_{pos} \geq n_{pos,min}) \& (\frac{n_{pos}}{n_{neg}} \geq f_{min})$ **then**
    accept k_i as true positive.
  else if $(n_{pos} < n_{pos,min}) \& (\frac{n_{pos}}{n_{neg}} < f_{min})$ **then**
    reject k_i
  else
    find the m keypoints $l_1 ... l_m$ closest to k_i
    if $l_1 ... l_m$ are all true positives **then**
      accept k_i as true positive
    else
      reject k_i
    end if
  end if
end for

function getKeypointList($candidate$):

  a) $list$ = all points in circular region around $candidate$ with $r = \sigma \cdot c, c = const$, σ = candidate scale
  b) $list$ = all point in circular region around $candidate$ with $r = c, c = const$
  c) $list$ = n spatially closest points to $candidate$
return $list$

Figure 1. Outline of the hysteresis-like candidate filtering algorithm.

4 Datasets

For training and testing, we recorded a dataset of 466 color images with PAL resolution. The set was recorded inside our intelligent house over different days and under varying lighting conditions. It contains images of 4 different persons wandering around inside our smart conference room and gesticulating. Note that we did not constrain the type of poses or gesticulations performed, that the persons appear in different distances to the cameras, and that they were allowed to move around freely in the camera's field of view. Fig. 2 shows some example images.

These images were segmented into hand and non-hand parts manually, where a small region around the perimeter of hands is also labelled as belonging to the hand to account for SIFT descriptors that describe typical hand regions, but lie outside the actual skin area. From this set, 145 images were randomly selected for testing. The remaining 321 images were taken for training of our classifier. The final database contains 200 000 SIFT descriptors for the background, and 8 700 for the foreground (i.e. the hands).

5 Experimental results

The ROC curve for NN matching using the full database of training examples is shown in Fig. 3, the variational parameter being the threshold t_{class} on the distance ratio s_{class}. It can be seen that, generally, the simple NN matching approach already yields satisfactory results. To get a high number of true positives, we allow for the accepted points to have a distance ratio $s_{class} > 1.0$, which means they are in fact more similar to some of the background examples. The point marked in Fig. 3 with 90% true positives[1] and 6.8% false positives corresponds to a threshold of 1.7. Due to the large number of key points that are identified by SIFT (typically 800-1600 per image), this results in a large number of false positives which typically lie on foreground objects (e.g. the person's body) not represented in the database. However, a considerable number of these will be discarded by our filtering algorithm.

[1] We assume that, for our application, a true positive rate lower than 85 to 90% will not be sufficient.

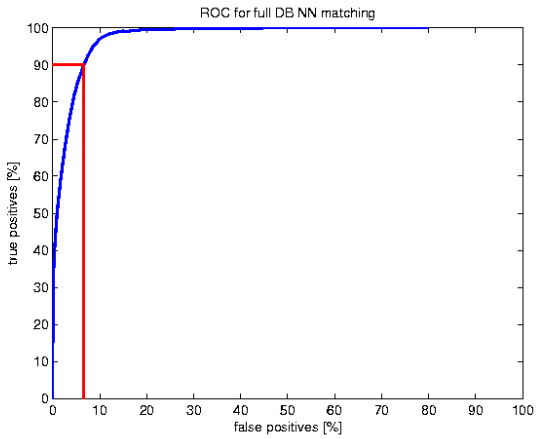

Figure 3. ROC curve for NN matching using complete training database.

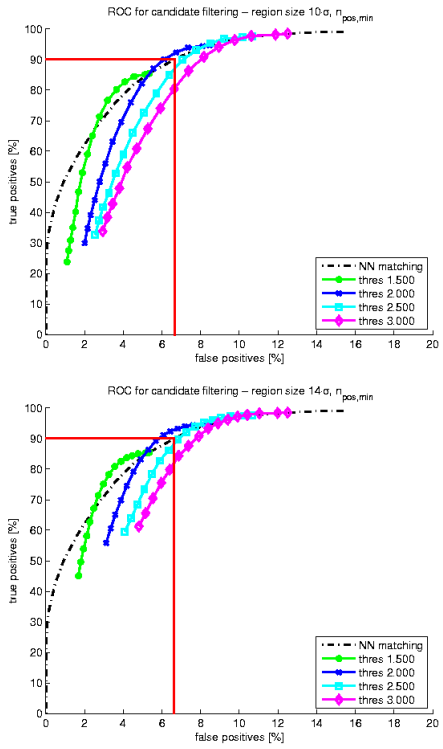

Figure 4. ROC curves for scale-dependent region-based filtering variating $n_{pos,min}$. Top: $c = 10$. Bottom: $c = 14$.

We now focus on the evaluation of our rejection criterion to reduce the number of false positives. In the following experiments, we evaluate the influence of the three parameters $n_{pos,min}$, f_{min} and m using different classification thresholds for the NN matching stage t_{class} and several region sizes. Because of space limitations, we can only show representative examples of our results.

Region-based filtering with scale-dependent region sizes: Since the SIFT procedure attaches to each keypoint the scale on which it was detected (which implicitly corresponds to the size of the object it describes), it seems natural to take the scale parameter as a hint for appropriate region size. Let σ_i be the scale parameter of keypoint k_i. We construct a circular region with radius $c \cdot \sigma_i$, where c is a constant, and determine the input list for the filtering algorithm from Fig. 1 from all keypoints inside that region.

Fig. 4 shows the ROC curves for two different values of c and different initial classification thresholds t_{class}, with the variational parameter being the required number of positive candidates in the region $n_{pos,min}$, ranging from 0 to 15. Note that the curves are scaled, since they start at the point defined by NN matching with threshold t_{class}.

As can be seen, the classification performance can be improved compared to the original NN matching (the dotted line). Good values for the initial classification threshold lie between 2.0 and 2.5. The same holds for the parameter f_{min}. The ROC curves for f_{min} ranging from 0 to 10 for a region radius of $10 \cdot \sigma$ are shown in Fig. 5.

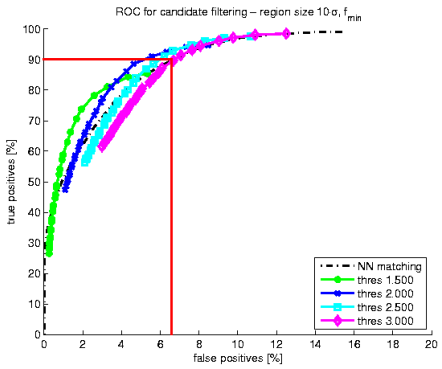

Figure 5. ROC curves for scale-dependent region-based filtering variating f_{min}. Region radius used here is $10 \cdot \sigma$.

Region-based filtering with fixed region sizes: The second approach we investigated uses circular regions of fixed sizes to determine the input list for our algorithm. Since the region size is not adjusted according to key point scale, this approach assumes that the sizes of the hands do not vary too strong in the input images. Fig. 6 shows the ROC curves for a region radius of 15 pixels when variating $n_{pos,min}$ (top) and f_{min} (bottom).

There seems to be no negative effect using the fixed region size compared to the scale-dependent approach. In fact, the results are better. The room the images were recorded in is quite small, and so the assumption that hand sizes do not vary strongly holds for most cases. Given a different scenario, a negative effect should be expected, but this is still to be investigated.

Filtering based on k nearest neighbors: Our last variant for candidate filtering does not use explicit regions, instead we generate the input list using the k spatially closest key points to the candidate. Fig. 7 shows the ROC curves for this approach. The results are slighly better than for the other 2 methods since we do not make assumptions on appropriate region sizes, but evaluate the same number of neighboring points for each candidate.

Second stage of filtering: The parameter m: Having evaluated the first stage of our candidate filtering algorithm with all three approaches for defining the local neighborhood which is used to "judge" the candidate key point, we can state that all three approaches work, in the sense that they considerably reduce the number of false positives while retaining a high true positive rate. Since there are no substantial differences in behavior, we will not favor one particular approach, but keep evaluating the second filtering stage for all three of them.

Above, we stated that we implemented a "hysteresis-like" rejection criterion. Looking again at the algorithm from Fig. 1, we see that the criteria evaluated so far – the minimum required number of positives $n_{pos,min}$ and the minimum ratio of positives and negatives f_{min} – both belong to the first stage and are applied simultaneously. However, failing one of them is not sufficient to reject a candi-

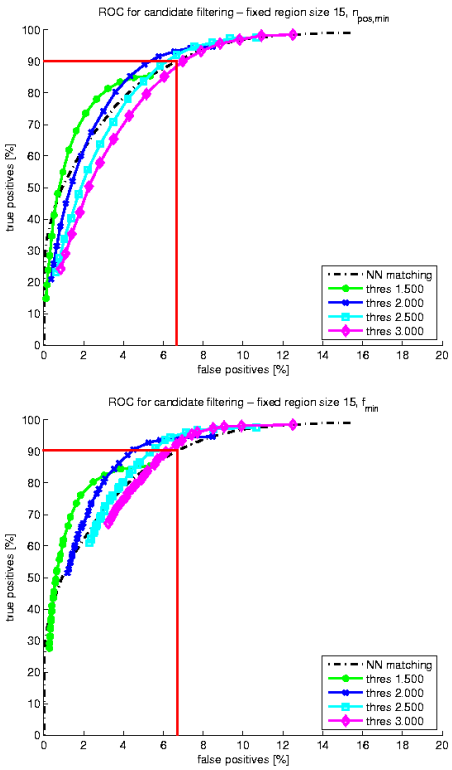

Figure 6. ROC curves for fixed-region based filtering. Top: Variating $n_{pos,min}$. Bottom: Variating f_{min}. The region radius used here is 15.

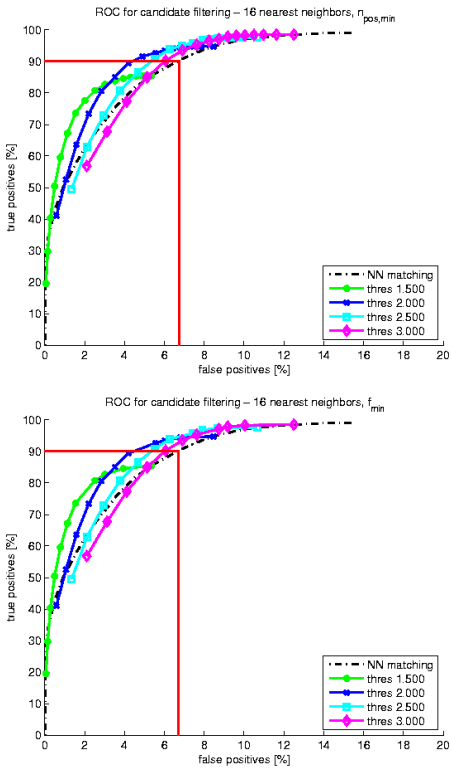

Figure 7. ROC curves for candidate filtering using 16 nearest neighbors. Top: Variating $n_{pos,min}$. Bottom: Variating f_{min}.

date key point. Instead, we only reject key points that fail on *both* conditions, and we only accept those that satisfy both. The remaining, which pass one criterion, but fail on the other, are further evaluated in a second stage.

Here, we look at the m spatially closest neighbors of the candidate point. We only accept it as true positive if *all* m neighbors are true positives, which means they must all have passed both conditions in the first stage. Setting m to a high value will eliminate "isolated" positives, but will also tend to discard key points at the margins of positive clusters. Figure 8 shows the ROC curves for different parameter combinations, variating m from 0 to 10. The plot has been scaled for better recognizability. Note that these curves have two fixed ends that are defined by the outcome of the first filtering stage: The starting point corresponds to the complete set of candidates that passed one of the two initial conditions, the end point to the number of candidates that passed both conditions. It can be seen that this is a pretty strong criterion, since for all values of $m > 0$, a certain portion of true positives is rejected. However, the effect on false positives is stronger. Since most false positives are already rejected for $m = 1$, and higher values for m will only discard more true positives, we will only take into consideration values of 1 and 2 for m for the evaluation of our complete system.

Evaluation of the complete system: Table 1 shows the results of a few example runs using parameter sets that seem reasonable based on our evaluation results of the different stages. For almost all parameter combinations, our filtering approach achieves a substantial reduction of the number of false positives while retaining true positive rates only slightly lower than in the initial NN matching stage. The best combinations reduce the number of false positives by one half, while only dropping around 5% of the true positives. This is an acceptable tradeoff, since our required rate of approximately 90% true positives can still be achieved in most cases.

Of our proposed variants for local neighborhood calculation, k nearest neighbors performs best. Looking at entry 5 in Table 1 and comparing with the "working point" marked for the straightforward NN matching scheme in Fig. 3, we see that for an equal true positive rate, the false positive rate is reduced by approximately 36% – which corresponds to more than 30 false detections per image for our test set. So, our candidate filtering algorithm improves the quality of our hand detection approach considerably.

6 Conclusion and Outlook

We presented an approach to hand detection using SIFT as one variant of a scale-invariant region descriptor. To reduce the number of false positives, we successfully utilised a two-stage, hysteresis-like filtering algorithm. Our results on a large set of images from a realistic indoor scene show

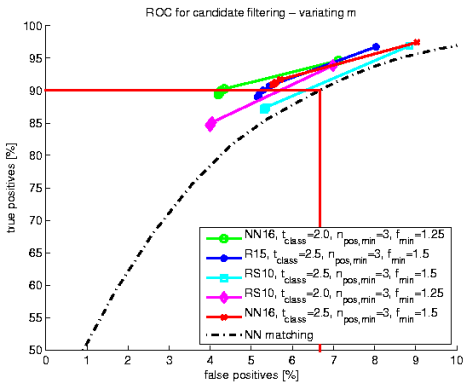

Figure 8. ROC curves for candidate filtering variating m with different parameter sets. NN16: 16 nearest neighbors. R15: fixed region size 15. RS10: region size $10 \cdot \sigma$.

	region	t_{class}	n	f_{min}	m	% TP (ΔNN)	% FP (ΔNN)
1	RS10	2.0	3	1.25	1	85.08 (-10.2)	4.04 (-52.3)
2	RS10	2.5	3	1.00	2	91.55 (-6.2)	6.32 (-40.8)
3	R15	2.0	4	1.25	1	88,18 (-6.9)	4.05 (-52.2)
4	R15	2.5	3	1.00	2	93.07 (-4.6)	6.00 (-43.8)
5	NN16	2.0	4	1.25	1	90.26 (-4.7)	4.34 (-48.8)
6	NN16	2.5	3	1.50	2	91.30 (-6.4)	5.61 (-47.4)

Table 1. Some example results using the complete filtering algorithm. $n_{pos,min}$ abbreviated to n. RS10: region size $10 \cdot \sigma$. R15: fixed region size 15. NN16: 16 nearest neighbors. The values in brackets give the relative drop for the true and false positive rates compared to NN matching with the same threshold t_{class}.

that this approach is very promising. However, so far it is not suitable for real-time applications, mainly due to the nearest-neighbor matching on a large descriptor database. We will investigate methods for database size reduction and accelleration of the matching scheme, as well as different classification approaches, in our future work.

Another issue concerns the features used: The SIFT descriptors are specifically designed for the tasks of image registration and rigid object detection, and therefore concentrate on finding local grey value distributions that are nearly identical to the ones extracted from the training images. They seem to lack the ability to cope with smooth deformations of strongly articulated objects, i.e. they do not generalize well. So, while the keypoint detection stage works fine – several characteristic points on hands are detected in almost all images – we will investigate different approaches for local feature representation. To conclude, our framework does not rely on SIFT being the feature extraction routine. Thus, any salient keypoint detection and representation scheme may be integrated easily, which makes the approach attractive for related applications.

References

[1] FINCA Project Homepage: http://www.finca.irf.de

[2] N. Hofemann, J.Fritsch and G.Sagerer, Recognition of deictic gestures with context, *Proc. 26th DAGM Symposium, LNCS*, Vol. 3175, Springer, Heidelberg, Germany, 2004, pp. 334-341.

[3] R. Lockton and A.W. Fitzgibbon, Real-time gesture recognition using deterministic boosting, *Proc. British Machine Vision Conference*, Cardiff, UK, 2002, pp. 817-826.

[4] K. Nickel and R. Stiefelhagen, Real-time person tracking and pointing gesture recognition for human-robot interaction, *Computer Vision in Human Computer Interaction, ECCV Workshop on HCI 2004, LNCS*, Vol. 3058, Springer, Prague, Czech Republic, 2004, pp. 28-38

[5] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Kauai, USA, 2001, pp. 511-518.

[6] T.F. Cootes and C.J. Taylor, Statistical models of appearance for medical image analysis and computer vision, *Image Processing - Proceedings of SPIE* , Volume 4322, 2001, pp. 238-248.

[7] J. Triesch and C. von der Malsburg, Classification of hand postures against complex backgrounds using elastic graph matching, *Image and Vision Computing*, Vol. 20, Elsevier Science, 2002, pp. 937-943.

[8] C. Harris and M. Stephens, A combined corner and edge detector, *Proc. of the Alvey Vision Conference*, 1988, pp. 147-151.

[9] T. Kadir and M. Brady, Scale, saliency and image description, *Int. Journal of Computer Vision*, Volume 45(2), 2001, pp. 83-105.

[10] H. Bay, T. Tuytelaars and L. van Gool, SURF: Speeded up robust features, *Proc. 9th European Conf. on Computer Vision, LNCS*, Vol. 3951, Springer, Graz, Austria, 2006, pp. 404-417.

[11] D.G. Lowe, Distinctive image features from scale-invariant keypoints, *Int. Journal of Computer Vision*, Volume 60, 2004, pp. 91-110.

[12] J. Liu and R. Hubbold, Automatic camera calibration and scene reconstruction with scale-invariant features, *Proc. 2nd Int. Symposium on Visual Computing, LNCS*, Vol. 4291, Springer, Lake Tahoe, USA, 2006, pp. 558-568.

[13] J. Kosecka and F. Li, Vision based topological markov localization, *Proc. IEEE Int. Conf. on Robotics and Automation*, Vol.2, New Orleans, USA, 2004, pp. 1481-1486.

[14] http://vision.ucla.edu/ vedaldi/code/sift/sift.html