

On Matching Interest Regions Using Local Descriptors - Can an Information Theoretic Approach Help?

Zoran Zivkovic Ben Kröse
Intelligent and Autonomous Systems Group
University of Amsterdam
The Netherlands
{zivkovic,krose}@science.uva.nl

Abstract

This paper shows that the common task of interest region matching using local descriptors can be improved using a new similarity measure. The similarity measure is motivated by the information theoretic image alignment that maximize mutual information between images. A property of the mutual information metric is that it does not only depend on how similar the signals are but also how complex they are. We present how similar logic can be applied to the standard SIFT descriptor. The results show improvement at almost no additional computational costs.

1 Introduction

There are many successful image analysis approaches for object recognition [6], 3D scene reconstruction [4] and other image analysis tasks that start with extracting interest regions from images. These approaches typically involve the following steps:

- **Detecting the interest regions.** An interest region detector defines a saliency measure and the interest regions are then detected by looking for the local maxima of the saliency measure across the image positions, across different sizes of the region and sometimes also across affine transformations of the region. The idea of checking different region sizes is to be able to detect the same region even if the region is present at different scales in different images. This leads to so called scale invariant detection. Checking also affine transformation of the region should make the detection affine invariant. The final result of interest region detection is a set of circular regions or elliptical regions if we also check the affine transformations. Years of experience produced a number of well behaved saliency measures and efficient methods for extracting the local maxima of the saliency measures. Various interest region detectors are analyzed and compared in [10] and [8]. A general conclusion from the empirical study [8] is that most of the modern interest point detectors give similar final matching results.

- **Computing the local image descriptors.** The goal is to describe each interest region by a descriptor vector computed from the local image values. In order to be able to easily compare the regions, the descriptor should be invariant to certain geometric transformations (scaling, rotations,...) and robust to some typical image changes (light changes, etc.). A good overview and comparison of various local image descriptors are given in [7]. In this study the SIFT descriptor [6] was shown to perform the best.
- **Matching the interest regions using the local image descriptors.** The regions are matched between pairs of images or between an image and a database of images. Some simple matching strategies are discussed in [7]. A similarity measure is defined between the region descriptor vectors in order to decide which pairs of points are the most likely matches. The measure should be fast and easy to compute since the matching usually involves computing the similarity for many of the possible region pairs. The similarity measure is commonly based on simple Euclidean distance of the descriptor vectors. There is no much other work on other similarity measures for the purpose of interest region matching.

This paper shows that the interest region matching using local descriptors can be improved using a new similarity measure. The new similarity measure is motivated by the work on information theoretic image alignment that maximize mutual information between images [12]. In Section 2 of the paper we describe the standard likelihood based similarity measures that lead to Euclidean distance based similarity. A general discussion about likelihood metric for image matching is given in [11]. Here we will study the problems that are specific for the task of fast interest point matching. We also demonstrate how this problems are usually approached by using the SIFT descriptor as an example. In Section 3 we discuss the mutual information based similarity metric and the difficulties of applying such a scheme on our problem. We show that the mutual information metric can be approximated by the standard likelihood metric that is penalized by a measure of complexity of the matching interest regions. In Section 4 we show how this fact can be applied in a simple way to the often used SIFT descriptor. We will use the standard Euclidean norm and penalize it so that the complex descriptors are preferred by the new measure. This will introduce a new similarity measure that is almost as easy to compute as the standard Euclidean based similarity. In Section 5 we show that the new measure greatly outperforms the Euclidean distance with almost no additional computational costs. A number of questions raised by these results are discussed in Section 6.

2 Likelihood similarity measure

The relation between the image values $v(x)$ from an interest region R with the corresponding image values $u(y)$ from the matching region from another image can be written in the following way:

$$v(x) = F(u(T(x)), \theta_{ext}) + \eta \quad (1)$$

The transformation T relates the local image coordinates x from one image to the local image coordinates of the other image $y = T(x)$ or inversely $x = T^{-1}(y)$. The transformation tells us which point from one image region corresponds to which point in the other

image. It also transforms the region R to its corresponding region $T(R)$. Since the interest regions are usually small parts of the image, simple affine transformation T is usually used as the transformation between the corresponding regions. The function F relates the corresponding image values between the 2 images and η is a random variable that models the noise. The external parameters θ_{ext} , that present for example light changes, motion blur, etc., are usually unknown and the following simplified model is often used:

$$v(x) \approx u(T(x)) + \eta \quad (2)$$

We can regard the image values $v(x)$ and $u(T(x))$ as random variables with associated probability density functions $p(v(x))$ and $p(u(T(x)))$. The probabilistic relation between the image values is expressed using their conditional probability density function $p(v(x)|u(T(x)))$. For example if the noise $\eta = N(0, V)$ is zero a mean Gaussian with covariance matrix V we have $p(v(x)|u(T(x))) = N(u(T(x)), V)$ or alternatively $p(u(y)|v(T^{-1}(y))) = N(v(T^{-1}(y)), V)$.

The discrete image points from the region R are denoted by x_1, \dots, x_N . Image values $v_i = v(x_i)$ and the corresponding image values from the other image $u_i = u(y_i) = u(T(x_i))$ can be considered as random samples. The vectors $\mathbf{v} = [v_1, \dots, v_N]^T$ and $\mathbf{u} = [u_1, \dots, u_N]^T$ are simple local descriptors of the regions. If we assume that the samples are independent the log-likelihood of the values from one region \mathbf{v} given the values \mathbf{u} from the other region and the alignment T is:

$$L(\mathbf{v}|\mathbf{u}, T) = \sum_{i=1}^N \log p(v(x_i)|u_i(T(x_i))) \quad (3)$$

The likelihood can be used as similarity measure between the regions. It is a symmetric measure since it is possible to show that $L(\mathbf{v}|\mathbf{u}, T) = L(\mathbf{u}|\mathbf{v}, T^{-1})$. We will use the following notation to denote this common similarity measure:

$$S_L(\mathbf{u}, \mathbf{v}, T) = L(\mathbf{v}|\mathbf{u}, T) = L(\mathbf{u}|\mathbf{v}, T^{-1}) \quad (4)$$

If we assume Gaussian noise $\eta = N(0, V)$ and diagonal matrix $V = \sigma^2 I$ we get the negative of the simple Euclidean distance:

$$S_{euclidian}(\mathbf{u}, \mathbf{v}, T) \sim -\frac{1}{\sigma^2} \sum_{i=1}^N (v(x_i) - u(T(x_i)))^2 \quad (5)$$

Note that the likelihood measure directly depends on the noise model η . In this paper we use the Gaussian model. However, other noise models might often be better as discussed in [11].

There are two other issues that should be considered when the likelihood similarity measure is used for the region matching:

External factors. The external factors θ_{ext} in (1) are unknown and disregarded in (2) but they could have a big influence on the values. Therefore the v_i and u_i from above are often not pure image values which are sensitive to such influences but some function of image values that are more robust. For example image gradient is often used since it is invariant to brightness changes. In this paper we will analyze in Section 4 the SIFT descriptor [6] that is based on the image gradient direction.

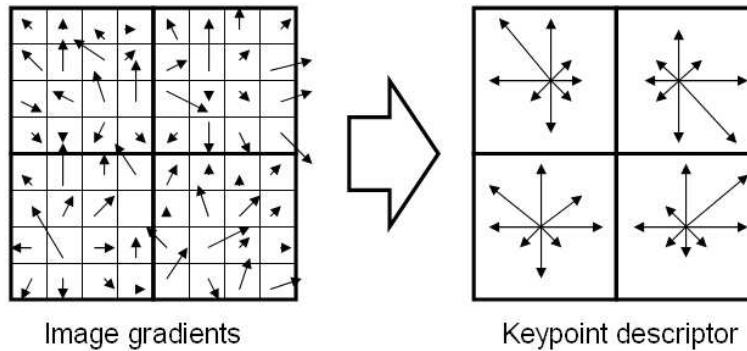


Figure 1: SIFT feature descriptor. The original 8x8 image grid of the interest region (on the left) is divided into 4 (2x2) sub-regions and for each of the sub-regions histogram of the image gradient orientations is calculated. For our results we used 16x16 original sample grid and 16 (4x4) subregions. See [6] for details.

Region alignment. The likelihood is highly sensitive to the alignment T . Maximizing the likelihood with respect to T is a standard image alignment technique. However matching interest regions would require the alignment between each pair of regions which is computationally expensive. Therefore a solution is to make the descriptor values v_i and u_i robust to some transformations. If we extract elliptical affine region we could first rescale the region to become circular. For the circular regions we then need the rotation that aligns the regions. Some approaches discussed in [7] try to make a rotation invariant descriptor but the distinctiveness of the descriptor is reduced. The approach that was performing the best is described in [6]. The local image gradients are used to determine the main orientation for each region. The main orientation is used to align the regions. See [6] for details. Anyhow, the scaling, the main orientation and often also the position of the point are coarse estimates of the actual transformation that aligns the regions. The descriptors \mathbf{v} and \mathbf{u} should be robust to such small deviations. Typically, the regions are divided into subregions and some histogram-like measure per subregion is used as descriptor. For example, the SIFT descriptor is composed of the gradient orientation histograms of the subregions as presented in Figure 1.

3 Mutual information similarity measure

The entropy of a random variable is given by:

$$H(v) = - \int p(v) \log p(v) dv \quad (6)$$

and the joint entropy of two random variables is:

$$H(v, u) = - \int \int p(v, u) \log p(v, u) dv du \quad (7)$$

where $p(v, u)$ is the joint probability density function. The entropy can be seen as a measure of complexity of a random variable or alternatively a measure of the amount of

information that the variable carries. The mutual information between two descriptors that are regarded as random signals is defined by:

$$S_{MI}(\mathbf{u}, \mathbf{v}, T) = I(v(x), u(T(x))) = H(v(x)) + H(u(T(x))) - H(v(x), u(T(x))) \quad (8)$$

The mutual information is an information theoretic measure that describes how much two random signals have in common. It takes values between 0 and 1. The joint entropy, the last term in (8), will be low if the two signals are related. However, if the two signals are simple then the first two terms will also be low. For example for two constant signals the entropy will be zero since the signals do not carry any information. A nice overview of different similarity measures and the relation with the mutual information is given in [12].

A standard problem when using the mutual information is the estimation of the underlying density functions. The joint density $p(v(x), u(T(x)))$ is most difficult to estimate since it has 2 times more dimensions than $p(v(x))$ and $p(u(T(x)))$. We can consider the values v_i and u_i that we observed from the images as random samples. In contrast to [12] where whole images are used, the typical number of samples for the interest regions is much smaller and estimates of $p(v(x), u(T(x)))$ will be of low quality. See [9] for some properties of the small sample estimates of the entropy. Furthermore, the region matching would require estimation of $p(v(x), u(T(x)))$ for each pair of regions and that would be computationally expensive.

After some manipulation [1] we can rewrite (8) as:

$$I(v(x), u(T(x))) = \int \int p(v(x), u(T(x))) \log p(v(x)|u(T(x))) + H(v(x)) \quad (9)$$

We often have some reasonable model for $p(v(x)|u(T(x)))$, for example Gaussian as in the previous section. Note that this is not the case in some medical applications which are the main topic in [12]. Since the values v_i and u_i that we observed from the images can be regarded as random samples from $p(v(x), u(T(x)))$ we can approximate the second term from (9) by its sample estimate:

$$I(v(x), u(T(x))) \approx \frac{1}{N} \sum_i^N \log p(u|v) + H(v(x)) = \frac{1}{N} S_L(\mathbf{u}, \mathbf{v}, T) + H(v(x)) \quad (10)$$

Mutual information is symmetric similarity measure but the approximation (10) depends on the order. By changing the order we can get $\frac{1}{N} S_L(\mathbf{u}, \mathbf{v}, T) + H(u(T(x)))$. In order to correct for this bias we could write:

$$S_{MI}(\mathbf{u}, \mathbf{v}, T) \approx \frac{1}{N} S_L(\mathbf{u}, \mathbf{v}, T) + \frac{1}{2} (H(v(x)) + H(u(T(x)))) \quad (11)$$

The above equation describes the link between the likelihood similarity measure and the mutual information measure. It follows that the mutual information can be approximated by the likelihood measure with an additional term that takes into account the complexity of the signals. The complexities of the signals can be computed offline. This also means that we could use some simple to compute likelihood based similarity measure, for example Euclidean based, and get an estimate of the mutual information measure at almost no additional computational costs.

Note that the approximation (11) depends on the correct conditional model $p(v(x)|u(T(x)))$. Given a model the parameters of the model should also be estimated. For example for the simple Gaussian (5) we need to estimate the standard deviation σ . If we do not have a correct model and/or we do not know the correct parameters of the model we can use some approximate model $q(v(x)|u(T(x)))$. It can be shown using Kullback-Liebler lower bound [1] that:

$$I(v(x), u(T(x))) \geq \int \int p(v(x), u(T(x))) \log q(v(x)|u(T(x))) + H(v(x)) \quad (12)$$

This means that the approximate solution using some model $q(v(x)|u(T(x)))$ for the likelihood similarity S_L actually approximates a lower bound of the mutual entropy. For example for Euclidean norm with unknown σ we could write:

$$S_{MI}(\mathbf{u}, \mathbf{v}, T) \approx \frac{\lambda}{N} S_{euclidian} + \frac{1}{2} (H(v(x)) + H(u(T(x)))) \quad (13)$$

where $\lambda = 1/\sigma^2$. Even with incorrect λ we will have some approximation of a lower bound of the mutual information.

4 Comparing SIFT descriptors

We could apply the same principle to the SIFT descriptors. However, the SIFT descriptor is a vector with 128 values that lie between 0 and 255. Estimation of the local entropy of the descriptor using only these 128 values would be tricky and the estimate would be of a low quality. On the other hand we observe that the SIFT descriptor presents a histogram of local image orientations by itself. The 8-bin histograms of the orientations are made for each of the 16 subregions of the interest region. The orientations are weighed by the strength of the gradients. Low gradient orientations are discarded and the high gradients are clipped. See Section 2 and [6] for more details. If $\mathbf{v} = v_1, \dots, v_N$ is a SIFT descriptor and if we normalize it so that it values sum to 1 we can use

$$H_v = - \sum_i^N v_i \log v_i \quad (14)$$

as a local measure of complexity of the gradient orientations within the interest region. Although they are computed on different domains we will still combine the Euclidean distance between the descriptors $S_{euclidian}(\mathbf{v}, \mathbf{u})$ and the local gradient ordination complexity measures H_v and H_u as in (13). This could be motivated by the fact that the Euclidean distance between the descriptors can be seen as some approximation of the likelihood similarity measure of the gradient orientations. We used fixed empirically chosen $\lambda = 1/400$ in our experiments. Compared to the Euclidean similarity, this measure requires minimal additional computational costs to compute the H_v and H_u . Furthermore, the H_v and H_u can be computed during feature point extraction and saved. Computation time for H_v and H_u is negligible compared to the time needed for the feature point extraction.

5 Experiments

We will evaluate the new metric in the same way as in [7] and on the same data. We used the SIFT detector and descriptor extraction routines that were provided by the author [6].



Figure 2: Data sets. First and the last image form each data set are presented.

Data sets. In order to evaluate the performance of the new matching measure we used some of the data-sets from [7]. The 4 data-sets we used contain 6 images under different changing conditions: light changes, image blur, zoom+rotation and JPEG compression. The changes become more drastic from image 1 to the last image in each data set. The first and the last image from each data set are presented in Figure 2. See [7] for more details about the data sets.

Ground truth. We use the same ground truth as in [7]. The interest regions are transformed using homography transformations between the images that is supplied with the data sets. We assume that a match is correct if the error in the image area covered by corresponding regions is less than 50% of the region union.

Evaluation. Same as in [7] we present the results as graphs that show *recall* versus $1 - \textit{precision}$. Recall is the number of correct matching regions with respect to the total number of corresponding regions:

$$\textit{recall} = \frac{\#correctmatches}{\#correspondences} \quad (15)$$

The number of false matches relative to the total number of matches is represented by the 1-precision:

$$1 - \textit{precision} = \frac{\#falsematches}{\#correctmatches + \#falsematches} \quad (16)$$

As in [7] we determine the matches by checking for each pair of points if the similarity between them is above a given threshold. We change the value of the threshold to plot the as *recall* versus $1 - \textit{precision}$ graphs. A perfect descriptor would give recall 1 for any precision but in practice the recall will increase with decreasing the threshold. A horizontal curve means that good recall can be obtained with high precision but there will be very similar structures in the scene that are not distinguishable by the descriptor.

Results. The results for all 4 data sets are presented in Figure 3. We compared the new similarity measure with the Euclidean similarity measure. The new similarity measure leads to improvements and similar performance with almost horizontal line on the graphs for all data sets. It is interesting to see that for all cases we can always get more than 60% of the features at very low false-positive rate. The SIFT descriptor divides the interest regions in blocks and this seems to make it robust to the JPEG compression that also produces blocky structures. Therefore there is a small improvement using the new

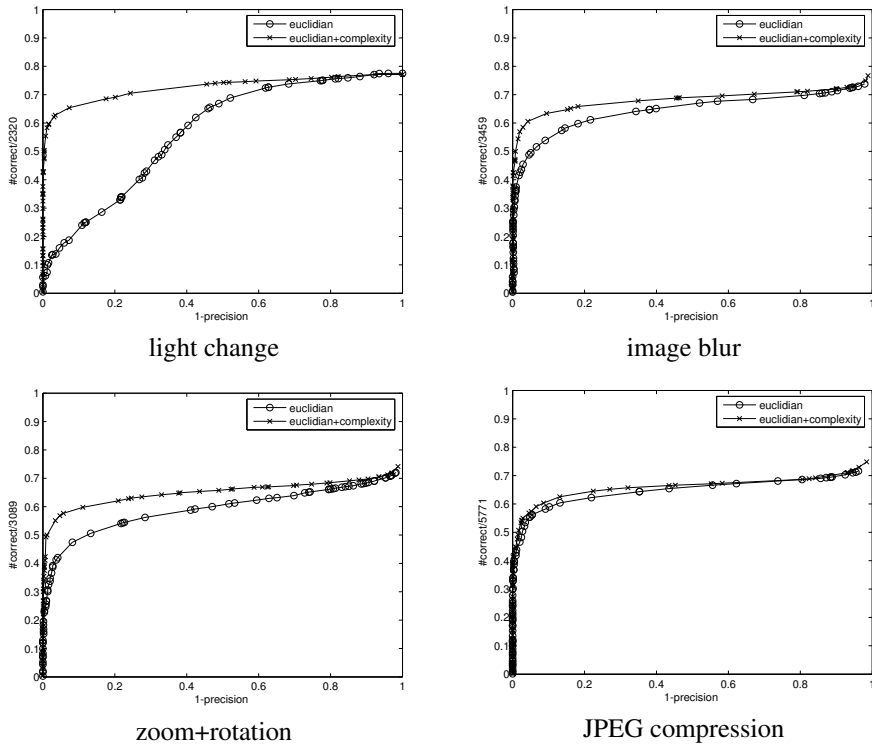


Figure 3: Results. The recall versus 1-precision graphs for each data set are presented.

measure on this data set. The improvement is the largest for the data set with strong light changes. It seems that SIFT descriptors become very similar for the very dark images and that the complexity term provides lot of useful information in such cases.

6 Conclusions and further work

We analyzed a number of similarity measures for the task of interest region matching using local descriptors. We have shown how the standard simple to compute Euclidean and other likelihood similarity measures can be extended by adding a term that considers the complexity of the matched signals in order to approximate the mutual information similarity measure. We present how similar logic can be applied to the standard SIFT descriptor. The results show improvement at almost no additional computational costs. The results could potentially be applied to other problems.

Typically region detection and matching are two separate processes. The new measure couples these two processes since the detection in general aims at detecting complex regions and the new measure takes the complexity also into account. There are interest region detectors that search for the regions with high entropy [3, 5]. The matching is the final goal of the interest region detection and it would be interesting to develop a top down approach for task of interest region detection and matching. Furthermore, adding other

problem specific knowledge should also be considered. For example in [2] the interest point selection is considered within a object classification framework.

Acknowledgments

The work described in this paper was conducted within the EU Project COGNIRON ("The Cognitive Companion") FP6-002020.

References

- [1] D. Barber and F.V. Agakov. The IM algorithm: A variational approach to information maximization. *NIPS*, 2003.
- [2] D. Gao and N. Vasconcelos. Discriminant saliency for visual recognition from cluttered scenes. *In Proceedings of Neural Information Processing Systems*, 2004.
- [3] S. Gilles. *Robust Description and Matching of Images*. University of Oxford, PhD Thesis, 1998.
- [4] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision, second edition*. Cambridge University Press, 2003.
- [5] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 2(45):83–105, 2001.
- [6] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2(60):91–110, 2004.
- [7] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Submitted to IEEE PAMI*, 2004.
- [8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Submitted to International Journal of Computer Vision*, 2004. Submitted in August 2004.
- [9] L Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1254, 2003.
- [10] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, 2(37):151–172, 2000.
- [11] N. Sebe, M.S. Lew, and D.P. Huijismans. Toward improved ranking metrics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1132–1141, 2000.
- [12] Paul A. Viola. *Alignment by Maximization of Mutual Information*. MIT, PhD Thesis, 1995.