Tutkimusraportti 111
Research Report 111

EXPERIMENTAL STUDY ON FAST 2D HOMOGRAPHY ESTIMATION FROM A FEW POINT CORRESPONDENCES

Joni-Kristian Kämäräinen and Pekka Paalanen

Lappeenranta University of Technology
Faculty of Technology Management
Department of Information Technology
Box 20
FIN-53851 Lappeenranta

ISBN 978-952-214-772-1 (paperback)
ISBN 978-952-214-773-8 (PDF)
ISSN 0783-8069

Lappeenranta 2009

Experimental Study on Fast 2D Homography Estimation from a Few Point Correspondences

Joni-Kristian Kamarainen[1,2], Pekka Paalanen[1]

[1]Machine Vision and Pattern Recognition Research Group
Lappeenranta University of Technology, Finland

[2]Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK

Abstract

Geometric transformation (homography) estimation from a few point correspondences is a necessary processing step in many feature-based object detection, localisation, registration and recognition methods. The homography estimation is only an intermediate and repeated step in often computationally very intensive methods and, therefore, fast estimation is an essential requirement. In this study we describe in detail and experimentally compare several estimation methods using simulated and real data and draw conclusions about which methods should be preferred. In addition, assuming planar models, we demonstrate how feature-based methods may in practise benefit from a more restricted homography (similarity, affinity) instead of allowing the genuine plane-geometric linear variation (projectivity). A more restricted transformation acts as a regulative factor providing necessary robustness when only a few correspondences are available.

1 Introduction

In recent state-of-the-art studies object detection, localisation, registration and recognition from 2D images are still solved as a 2D-to-2D problem despite the fact that these tasks actually belong to a 3D-to-2D problem domain (e.g. participants in the VOC2007 challenge [4]). This mistreat provides many insuperable problems, which have been recognised but only lately resulted to an increasing interest toward 3D-to-2D methods (e.g. [16]). 2D-to-2D methods still outperform 3D-to-2D methods and also in this study we directly contribute to the "traditional approach" by investigating the linear 2D geometry transformation (homography) estimation which is at the core of most matching methods. A similar problem, pose estimation, is however an intrinsic part of the 3D-to-2D methods as well. Our interest is to accurately estimate a homography from a 2D object model space to 2D image space and vice versa. This process is especially important for part-based methods where object parts are detected from an image, and a part-based object model is fitted to the detected parts by looking for the best match [5, 7, 10, 13, 14]. It is clear, that in methods requiring an exhaustive or random search the homography estimation must be very fast and succeed with only a few

extracted part (usually point) correspondences. Since we have faced this problem in our previous studies [3, 7, 13] we were motivated to study the homography estimation itself. In this study we introduce a set of well-known and some less known fast methods for estimation of homographies of various types (isometry, similarity, affinity and projectivity) and compare them using realistically generated simulation data and real data sets for human face detection. The experiments on the real data support the findings with the simulated data and therefore allows us to draw conclusions for methods preferable for estimating homographies of a specific type from a few point correspondences. Our data and the homography estimation methods are provided as Matlab source code at the project web-page: `http://www.it.lut.fi/project/homogr/`.

2 2D homography

Homography is a linear geometric transformation between two entities in N dimensions forming a group (two or more consecutive transformations can be expressed as a single transformation of the same type). 2-D geometric transformations are frequently used in computer vision and they are expressed in projective 3-space, i.e. in homogenous coordinates. The most important linear transformations with their degrees of freedom (dof) explained are:

1. Isometry transformation (isometry) (plane rotation and translation, 3 dof, $\boldsymbol{H}_{\mathrm{E}}$)
2. Similarity transformation (similarity) (+ orthogonal depth translation, i.e. isotropic scaling, 4 dof, $\boldsymbol{H}_{\mathrm{S}}$)
3. Affine transformation (affinity) (+ rotation in depth with weak perspective effect (shear), i.e. non-isotropic scaling, 6 dof, $\boldsymbol{H}_{\mathrm{A}}$)
4. Projective transformation (projectivity) (+ perspective effect, 8 dof, $\boldsymbol{H}_{\mathrm{P}}$)

It should be noted that $\boldsymbol{H}_{\mathrm{E}} \subset \boldsymbol{H}_{\mathrm{S}} \subset \boldsymbol{H}_{\mathrm{A}} \subset \boldsymbol{H}_{\mathrm{P}} = \mathbb{R}^{3\times 3}$The structure of the homography matrix is (notation from [8])

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{t} \\ \boldsymbol{v}^{\mathrm{T}} & v \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & t_{\mathrm{x}} \\ a_3 & a_4 & t_{\mathrm{y}} \\ v_1 & v_2 & v \end{bmatrix} \quad . \tag{1}$$

In addition to the above homographies, it is possible to define two more 2-D transformations, but these have been rarely used. It can be seen from the degrees of freedom between similarity and affinity, and affinity and projectivity, that there are some transformations "between" them. They can, for example, be defined as:

2.5. Similarity+ transformation (5 dof, $\boldsymbol{H}_{\mathrm{S}+}$)

3.5. Affinity+ transformation (7 dof, $\boldsymbol{H}_{\mathrm{A}+}$)

These transformations require parameters from the next level to be constrained (e.g. one of them fixed). Interpretations of these transformations depend on how the parameters are constrained. The constraints can, for example, correspond to some mapping

of rigid transformations of 3-D points to an image plane for a certain type of camera. Constraining the parameters makes the transformations "non-symmetric", and therefore, we must denote by $\boldsymbol{H}_{\mathrm{S}+}$ and $\boldsymbol{H}_{\mathrm{A}+}$ the transformations from a "model space" to "image space" and $\boldsymbol{H}_{\mathrm{S}\oplus}$ and $\boldsymbol{H}_{\mathrm{A}\oplus}$ vice versa. In our definition, the model space corresponds to some "aligned space" where the constraints hold (e.g. fixed parameters have their true values set). In the image space the parameters are affected by other transformation parameters. Now, knowing the transformation direction is important since the order of linear operations must be changed, and most importantly, only now it is possible to derive formula for estimating the transformation parameters. The two transformations are not subspaces of the complete homography operator space since they do not form a group.

2.1 Homography matrices and their parametrisation

In the following definitions we follow the notations in [8]. $\boldsymbol{x}$ is a homogenous 3-vector and $\boldsymbol{t}$ is an inhomogenous translation 2-vector.

Isometry is defined as

$$\boldsymbol{x}' = \boldsymbol{H}_{\mathrm{E}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \epsilon\cos\theta & -\sin\theta & t_{\mathrm{x}} \\ \epsilon\sin\theta & \cos\theta & t_{\mathrm{y}} \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{x} \tag{2}$$

where $\epsilon = 1$ if the isometry is orientation preserving (Euclidean transformation) and -1 if orientation reversing (mirroring). In computer vision the mirroring can be useful if, for example, an object has sensible meaning its visual presence mirrored, e.g. side view of a motor bike or car, or in computer graphics for producing reflections. Mirroring can be allowed in isometry and similarity, but it is not the typical case. Mirroring transformations do not form a group with themselves, but combined with non-mirroring transformations they do. If $\epsilon = \{1, -1\}$ is allowed, then the estimation may need an explicit usage of the both values.

The definition of the similarity transformation is (Euclidean transformation with scaling) [8]:

$$\boldsymbol{x}' = \boldsymbol{H}_{\mathrm{S}}\boldsymbol{x} = \begin{bmatrix} s\boldsymbol{R} & \boldsymbol{t} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_{\mathrm{x}} \\ s\sin\theta & s\cos\theta & t_{\mathrm{y}} \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{x} \tag{3}$$

where s is an isotropic scale (scaling factor).

Affinity is defined as

$$\boldsymbol{x}' = \boldsymbol{H}_{\mathrm{A}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{t} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & t_{\mathrm{x}} \\ a_{21} & a_{22} & t_{\mathrm{y}} \\ 0 & 0 & 1 \end{bmatrix} \boldsymbol{x} \tag{4}$$

where $\boldsymbol{A}$ is a non-singular matrix.

Projectivity is a general non-singular linear transformation of homogeneous coordinates defined as

$$\boldsymbol{x}' = \boldsymbol{H}_{\mathrm{P}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{t} \\ \boldsymbol{v}^{\mathrm{T}} & v \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} a_{11} & a_{12} & t_{\mathrm{x}} \\ a_{21} & a_{22} & t_{\mathrm{y}} \\ v_1 & v_2 & v \end{bmatrix} \boldsymbol{x} \tag{5}$$

containing nine elements, but only their ratio is significant (eight degrees of freedom).

2.2 Additional example: similarity+

Similarity transformation has 4 degrees of freedom and the next homography, affinity, 6. Similarity requires at least 2 points correspondences and affinity 3. There clearly is one more transformation between the two homographies, a transformation of 5 degrees of freedom. For this transformation we cannot use the similarity form in (3) but it is a "sub-transformation" of affinity in (4). Let us next formulate a naive example of constrainted affinity, similarity+, by simply fixing one of the affine transformation parameters reducing degrees of freedom to 5 dof.

The affinity has several different decompositions, one of the most interpretable being in [8], where the matrix $\boldsymbol{A}$ is decomposed to the standard rotation and non-isotropic scaling (λ_1, λ_2) applied at a certain "scaling angle" ϕ as

$$\boldsymbol{A} = \boldsymbol{R}(\theta)\boldsymbol{R}(-\phi)\boldsymbol{D}\boldsymbol{R}(\phi) \tag{6}$$

where

$$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \quad . \tag{7}$$

In our naive example we simply fix the scaling angle ϕ.

There appears a problem with the fixed angle: the points must always be first rotated to a predefined standard pose where the scaling angle has its true value before applying the scaling. We may fix, for example, that the points in the model space are in the standard pose. That makes the similarity+ transformation different depending whether your are transforming correspondence points from image space to the model space or from the model space to the image space. If the transformation is made from the model to the image space, then the scaling and rotation part $\boldsymbol{A}$ is

$$\boldsymbol{A}_+ = \boldsymbol{R}(\theta)\boldsymbol{R}(-\phi_{fix})\boldsymbol{D}\boldsymbol{R}(\phi_{fix}) \tag{8}$$

and if the transformation is made from image to model space

$$\boldsymbol{A}_\oplus = \boldsymbol{R}(-\phi_{fix})\boldsymbol{D}\boldsymbol{R}(\phi_{fix})\boldsymbol{R}(\theta) \quad . \tag{9}$$

Finally, utilising our assumptions the similarity+ transformation from the model space to the image space can be defined as

$$\boldsymbol{x}' = \boldsymbol{H}_{\mathrm{S}+}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{A}_+ & \boldsymbol{t} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \tag{10}$$

and from the image space to the model space as

$$\boldsymbol{x}' = \boldsymbol{H}_{\mathrm{S}\oplus}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{A}_\oplus & \boldsymbol{t} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \quad . \tag{11}$$

It is clear that these special transformations require special attention in computing and estimating their parameters as compared to the true homographies.

3 Fast—linear and nearly linear—homography estimation from point correspondences

In correspondence based estimation we utilise two sets of geometric entities with known correspondence labelling and related by an unknown homography. The simplest class of entities (not necessarily from the estimation point of view) are points. In that case, the homography is estimated by computationally solving the geometric relationship between the two point sets, $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ and $\{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \ldots, \boldsymbol{x}'_N\}$, where the point $\boldsymbol{x}_i$ in the first set corresponds to the point $\boldsymbol{x}'_i$ in the second set, i denoting the label. Throughout the text we reference to the first set as "model" or source and the second as "observation" or image or target, since in our experiments the first corresponds to known (annotated) landmarks derived from an object description and the second extracted (possibly noisy or even erroneous) local features extracted from an input image. The terms are interchangeable.

Our problem domain requires fast computation which restricts approaches to linear or some very simple iterative, "nearly linear", methods. In object and especially object category (e.g. faces, license plates) detection, localisation and recognition problems descriptions of objects (model) are not rigorous but stochastic in nature and, on the other hand, all measurements from an input image (observations) must also be considered noisy. Therefore, the only proper error measure would be the *re-projection error* which simultaneously estimates both the homography and the true point locations using statistical models of errors in the both spaces [8]. However, optimising re-projection error or even its simpler counterparts, such as error in observation image only or symmetric transfer error, require computationally intensive iterative processing which is not feasible in algorithms where the homography estimation is only a repeated intermediate part. Our problem is thus completely different as compared to, for example, camera calibration or accurate scene reconstruction where high accuracy is the essential goal. Next we discuss fast methods which can be computed rapidly and thus be parts of real-time systems.

3.1 Exact solution

We denote an estimate as an exact solution if it maps source points exactly to target points and is the unique solution. This, in practice, is possible only for the minimum required number of correspondences for a homography class. With less points the solution is not unique and with more points there is no guarantee that an exact solution exists. In some critical configurations, even with the minimum number of correspondences the solution may not be unique, but these configurations can be identified. The critical configurations may appear for any number of point correspondences. Estimation using the minimum number of correspondences is useful for many applications. In the following exact solutions are derived for several homography classes.

3.1.1 Isometry

A transformation representing the isometry between two point sets can be uniquely solved from two point correspondences, given point sets $A = \{(x_1, y_1), (x_2, y_2)\}$ and $B = \{(x'_1, y'_1), (x'_2, y'_2)\}$. The isometry transformation $B = \boldsymbol{H}_{\mathrm{E}}(A)$ can be exactly

solved by the following steps (point transforms used instead of coordinate system transforms to simplify notation).

One of the correspondences, e.g., (x_1, y_1) and (x'_1, y'_1), is selected as a pivot element. First, the point set A is translated by moving the pivot element to the origin

$$\boldsymbol{A}_{\mathrm{o}} = \begin{bmatrix} 1 & 0 & -x_1 \\ 0 & 1 & -y_1 \\ 0 & 0 & 1 \end{bmatrix} \quad . \tag{12}$$

Next, the points are rotated according to the angular difference $\hat{\theta} = \theta' - \theta$ to the same orientation as the points in the set B

$$\boldsymbol{R}(\hat{\theta}) = \begin{bmatrix} \cos\hat{\theta} & -\sin\hat{\theta} \\ \sin\hat{\theta} & \cos\hat{\theta} \end{bmatrix} \tag{13}$$

Finally, the points are translated by moving the pivot element to the pivot position in B

$$\boldsymbol{A}_{\mathrm{B}} = \begin{bmatrix} 1 & 0 & x'_1 \\ 0 & 1 & y'_1 \\ 0 & 0 & 1 \end{bmatrix} \quad . \tag{14}$$

The complete isometry transformation of the points can thus be computed from

$$\boldsymbol{x}_{\mathrm{B}} = \boldsymbol{H}_{\mathrm{E}}\boldsymbol{x}_{\mathrm{A}} = \boldsymbol{A}_{\mathrm{B}}\boldsymbol{R}_{\hat{\theta}}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{x}_{\mathrm{A}} \tag{15}$$

where the homogenous rotation matrix is used:

$$\boldsymbol{R}_{\hat{\theta}} = \begin{bmatrix} \boldsymbol{R}(\hat{\theta}) & \boldsymbol{0} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} \quad . \tag{16}$$

The rotation matrix can be formulated as follows. Let $(d_{\mathrm{x}}, d_{\mathrm{y}}) = (x_2, y_2) - (x_1, y_1)$ and $(d'_{\mathrm{x}}, d'_{\mathrm{y}}) = (x'_2, y'_2) - (x'_1, y'_1)$. With the angle θ defined as the angle between the x-axis and the direction of $(d_{\mathrm{x}}, d_{\mathrm{y}})$ it follows that $\cos\theta = \frac{d_{\mathrm{x}}}{r}$ and $\sin\theta = \frac{d_{\mathrm{y}}}{r}$, where $r = \|(d_{\mathrm{x}}, d_{\mathrm{y}})\|_2$, and θ' similarly. If the point sets are related strictly by an isometry, then $r = r'$. Using the identities

$$\begin{aligned} \cos(\theta' - \theta) &= \cos\theta'\cos\theta + \sin\theta'\sin\theta \\ \sin(\theta' - \theta) &= \sin\theta'\cos\theta - \cos\theta'\sin\theta \end{aligned} \tag{17}$$

the rotation matrix can be written as

$$\boldsymbol{R}(\hat{\theta}) = \frac{1}{rr'}\begin{bmatrix} d_{\mathrm{x}}d'_{\mathrm{x}} + d_{\mathrm{y}}d'_{\mathrm{y}} & -d_{\mathrm{x}}d'_{\mathrm{y}} + d'_{\mathrm{x}}d_{\mathrm{y}} \\ d_{\mathrm{x}}d'_{\mathrm{y}} - d'_{\mathrm{x}}d_{\mathrm{y}} & d_{\mathrm{x}}d'_{\mathrm{x}} + d_{\mathrm{y}}d'_{\mathrm{y}} \end{bmatrix} = \frac{1}{rr'}\begin{bmatrix} d'_{\mathrm{x}} & d'_{\mathrm{y}} \\ d'_{\mathrm{y}} & -d'_{\mathrm{x}} \end{bmatrix}\begin{bmatrix} d_{\mathrm{x}} & d_{\mathrm{y}} \\ d_{\mathrm{y}} & -d_{\mathrm{x}} \end{bmatrix} \quad . \tag{18}$$

Using equations (12), (14) and (18) the isometry transformation $\boldsymbol{H}_{\mathrm{E}} = \boldsymbol{A}_{\mathrm{B}}\boldsymbol{R}_{\hat{\theta}}\boldsymbol{A}_{\mathrm{o}}$ can be solved using two point correspondences. In the experiments this method is denoted **EX-ISO** (exact isometry).

3.1.2 Similarity

The decomposition of the similarity transformation is similar to the isometry in (15), but including also an isotropic scale factor, s, in a scaling matrix

$$\boldsymbol{S} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad . \tag{19}$$

Now, the decomposition for the similarity is

$$\boldsymbol{x}_{\mathrm{B}} = \boldsymbol{H}_{\mathrm{E}}\boldsymbol{x}_{\mathrm{A}} = \boldsymbol{A}_{\mathrm{B}}\boldsymbol{S}\boldsymbol{R}_{\hat{\theta}}\boldsymbol{A}_{\mathrm{o}}\boldsymbol{x}_{\mathrm{A}} \quad . \tag{20}$$

The exact solution for the similarity transformation is the same as for the isometry except that the scale factor is solved from

$$s = \frac{r'}{r} \quad . \tag{21}$$

This method is denoted **EX-SIM** (exact similarity).

3.1.3 Affinity

The exact solution for the affinity is straightforward in the matrix form

$$\begin{bmatrix} x'_1 & x'_2 & x'_3 \\ y'_1 & y'_2 & y'_3 \\ 1 & 1 & 1 \end{bmatrix} = \boldsymbol{X}' = \boldsymbol{H}_{\mathrm{A}}\boldsymbol{X} = \boldsymbol{H}_{\mathrm{A}} \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix} \Rightarrow \boldsymbol{H}_{\mathrm{A}} = \boldsymbol{X}'\boldsymbol{X}^{-1} \tag{22}$$

where the last row of $\boldsymbol{H}_{\mathrm{A}}$ is automatically enforced to $(0, 0, 1)$ and the other 6 elements are free. This produces a general non-singular (for a non-degenerate point configuration) linear transformation of inhomogeneous coordinates and a translation.

3.1.4 Projectivity

The exact projectivity (up to a non-zero scale factor) can be computed from 4 point correspondences, but is postponed until Sec. 3.3 since it is also a linear solution for more than 4 points.

3.1.5 Additional example: similarity+

In the following we demonstrate exact solution for the naive similarity+ transformation given in Sec. 2.2. The similarity+ transformation requires 3 point correspondences (two and a half actually) to be solved. Let A be the model space and B the image space. The exact solution for the similarity+ transformation begins from the fact that in each space the lengths of vectors defined by the three given points are related by the non-isotropic scaling ${}^{\mathrm{B}}\boldsymbol{S}_{\mathrm{A}} = \boldsymbol{R}(-\phi)\boldsymbol{D}\boldsymbol{R}(\phi)$. The similarity+ transform

$${}^{\mathrm{B}}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{R}(\theta)\,{}^{\mathrm{B}}\boldsymbol{S}_{\mathrm{A}} & {}^{\mathrm{B}}\boldsymbol{t}_{\mathrm{A}} \\ \boldsymbol{0}^{\mathrm{T}} & 1 \end{bmatrix} {}^{\mathrm{A}}\boldsymbol{x} \quad . \tag{23}$$

Once ${}^{\mathrm{B}}\boldsymbol{S}_{\mathrm{A}}$ is known, the problem reduces into an isometry estimation.

The unknowns in ${}^{\mathrm{B}}\boldsymbol{S}_{\mathrm{A}}$ are the elements λ_1, λ_2 of the diagonal matrix $\boldsymbol{D}$, ϕ is a known constant. For two unknowns, two constraints are needed. Choose one correspondence $i = 1$ as the pivot and denote the vectors from ${}^{\mathrm{A}}\boldsymbol{x}_i$ to ${}^{\mathrm{A}}\boldsymbol{x}_j | j = 2, 3$ as ${}^{\mathrm{A}}\boldsymbol{v}_j$. Define the rotation basis vectors

$$\boldsymbol{e}_1 = \begin{bmatrix} \cos\phi \\ \sin\phi \end{bmatrix} \text{ and } \boldsymbol{e}_2 = \begin{bmatrix} -\sin\phi \\ \cos\phi \end{bmatrix} \ , \tag{24}$$

so that $\boldsymbol{R}(\phi) = [\boldsymbol{e}_1 \ \boldsymbol{e}_2]$. The squared Euclidean length

$$\begin{aligned} \| {}^{\mathrm{B}}\boldsymbol{v}_j \|^2 &= \| \boldsymbol{D}\boldsymbol{R}(\phi) \, {}^{\mathrm{A}}\boldsymbol{v}_j \|^2 \\ &= \left[\lambda_1 (\boldsymbol{e}_1 \cdot {}^{\mathrm{A}}\boldsymbol{v}_j) \right]^2 + \left[\lambda_2 (\boldsymbol{e}_2 \cdot {}^{\mathrm{A}}\boldsymbol{v}_j) \right]^2 \\ &= \lambda_1^2 u_j^2 + \lambda_2^2 v_j^2 \ , \end{aligned} \tag{25}$$

where $u_j = \boldsymbol{e}_1 \cdot {}^{\mathrm{A}}\boldsymbol{v}_j$ and $v_j = \boldsymbol{e}_2 \cdot {}^{\mathrm{A}}\boldsymbol{v}_j$. There is a pair of equations with $j = 2, 3$:

$$\begin{cases} \lambda_1^2 u_2^2 + \lambda_2^2 v_2^2 &= \| {}^{\mathrm{B}}\boldsymbol{v}_2 \|^2 \\ \lambda_1^2 u_3^2 + \lambda_2^2 v_3^2 &= \| {}^{\mathrm{B}}\boldsymbol{v}_3 \|^2 \end{cases} \ , \tag{26}$$

which can be written in matrix form as

$$\begin{bmatrix} u_2^2 & v_2^2 \\ u_3^2 & v_3^2 \end{bmatrix} \begin{bmatrix} \lambda_1^2 \\ \lambda_2^2 \end{bmatrix} = \begin{bmatrix} \| {}^{\mathrm{B}}\boldsymbol{v}_2 \|^2 \\ \| {}^{\mathrm{B}}\boldsymbol{v}_3 \|^2 \end{bmatrix} \ . \tag{27}$$

The solution for λ_1, λ_2 is obvious by multiplying the matrix equation with the inverse matrix from the left. Positive roots must be chosen for the lambdas to avoid mirroring.

If the 2D point are not in a degenerate configuration (on a line or at a single point), the uv-matrix is invertible. This follows from the fact, that the (u, v) vectors must then span the 2D plane.

3.2 Iterative Exact Solution: Random Sampling

If there are more than the minimum required number of correspondences available it is possible to operate in the spirit of the random sample consensus (RANSAC) [6]. Then the solution is estimated by randomly selecting the minimum number of correspondences and utilising an exact solution method. The RANSAC solution is more robust to noise and also to outliers if inlier detection is used. The best transformation can be selected according to some error measure, e.g. transfer error or symmetric transfer error. We may define at least two different kinds of iterative exact methods:

1. Iterative cross-selection
2. Iterative weighted averaging

The first method is a straightforward generalisation of the exact formula over more than the minimum number of points: after N samples the homography producing the smallest error is selected. If prior information of correspondence uncertainties is available an optimal selection of N and inlier threshold can be established.

The second approach corresponds to the first, but instead of using the single best estimate it should algebrailly or statistically "weight" the different estimates based on their confidence. In our experiments only the first method was implemented due to its simplicity: only N needs to be defined and the transfer error was used as the error measure (**RANSAM-***).

3.3 Direct linear transform (DLT)

The problem of all methods based on the exact estimation formula is that solutions are always derived from a minimum set of correspondences. It makes more sense that the solution should combine information from all point correspondences. One of the most widely used method using an overdetermined set of correspondences is the direct linear transform (DLT), which forms a linear system and computes the null-space solution by the singular value decomposition (SVD) [8]. According to Hartley and Zisserman [8], the solution can be derived from the vector cross product of point correspondences in homogeneous coordinates, $\boldsymbol{x}_i = (x_i, y_i, w_i)^{\mathrm{T}}$ and $\boldsymbol{x}'_i = (x'_i, y'_i, w'_i)^{\mathrm{T}}$, as

$$\boldsymbol{x}'_i \times \boldsymbol{H}\boldsymbol{x}_i = \boldsymbol{0} \ . \tag{28}$$

The cross product can be written as a linear system

$$\begin{bmatrix} \boldsymbol{0}^{\mathrm{T}} & -w'_i\boldsymbol{x}_i^{\mathrm{T}} & y'_i\boldsymbol{x}_i^{\mathrm{T}} \\ w'_i\boldsymbol{x}_i^{\mathrm{T}} & \boldsymbol{0}^{\mathrm{T}} & -x'_i\boldsymbol{x}_i^{\mathrm{T}} \\ -y'_i\boldsymbol{x}_i^{\mathrm{T}} & x'_i\boldsymbol{x}_i^{\mathrm{T}} & \boldsymbol{0}^{\mathrm{T}} \end{bmatrix} \begin{pmatrix} \boldsymbol{h}^1 \\ \boldsymbol{h}^2 \\ \boldsymbol{h}^3 \end{pmatrix} = \boldsymbol{0} \tag{29}$$

where $\boldsymbol{h}^{i\mathrm{T}}$ are the corresponding rows of the homography matrix. Only two of the rows are independent and therefore one of them can be discarded and the remaining ones written in an open form as (the first and second row)

$$\begin{bmatrix} 0 & 0 & 0 & -w'_ix_i & -w'_iy_i & -w'_iw_i & y'_ix_i & y'_iy_i & y'_iw_i \\ w'_ix_i & w'_iy_i & w'_iw_i & 0 & 0 & 0 & -x'_ix_i & -x'_iy_i & -x'_iw_i \end{bmatrix} \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \end{pmatrix} = \boldsymbol{0} \ . \tag{30}$$

It is clear that this system can provide a proper solution only for projectivity in (5) (up to a non-zero scale factor) since it contains 9 free parameters. For an arbitrary number of correspondences the system only very rarely leads to an exact solution in the presence of noise. An optimisation method for solving the linear system is needed. The exact solution would require four and a half 2-D point correspondences. For an overdetermined case a singular vector corresponding to the smallest singular value in the SVD provides a solution with the minimal algebraic error (sum of squared residuals in target space). The less noisy and more well-configured the correspondences are, the more stable and accurate is the estimate. The most well-known methods based on this principle are the "Golden standard" method from [8] ([17]) and the methods fixing $H(3,3) = 1$ and utilising least-squares [2] or pseudo-inverse [1] ([9]). These methods solve the projective transformation matrix and are called as direct linear transformation (DLT) methods. For unnormalised data the two rather historical methods make sense, but in general, only the golden standard method should be utilised. The normalisation procedure is discussed later.

For a large number of points these methods are practically almost as good as the more dedicated non-linear methods [8] and due to their computational efficiency these methods are the most popular in many computer vision related problems, such as in providing an initial estimate in camera calibration.

3.3.1 Formulations for more restricted transformations

For the more restricted homographies than the projective transformation, the linear system in (30) can be further restricted. For instance, by using the following equalities derived from the isometry in (2)

$$\boldsymbol{H}_{\mathrm{E}} \Leftrightarrow \begin{pmatrix} \boldsymbol{h}^1 \\ \boldsymbol{h}^2 \\ \boldsymbol{h}^3 \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \end{pmatrix} = \begin{pmatrix} \cos\theta \\ -\sin\theta \\ t_x \\ \sin\theta \\ \cos\theta \\ t_y \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} a \\ -b \\ c \\ b \\ a \\ d \\ 0 \\ 0 \\ 1 \end{pmatrix} \tag{31}$$

the system can be simplified to

$$\begin{bmatrix} -w_i' y_i & -w_i' x_i & 0 & -w_i' w_i \\ w_i' x_i & -w_i' y_i & w_i' w_i & 0 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = \begin{pmatrix} -y_i' w_i \\ x_i' w_i \end{pmatrix} \tag{32}$$

which can be further organised into a form solvable by the SVD method. From the solution the transformation matrix (isometry) can be constructed as

$$\left.\begin{matrix} \boldsymbol{H}_{\mathrm{E}} \\ \boldsymbol{H}_{\mathrm{S}} \end{matrix}\right\} = \begin{bmatrix} a & -b & c \\ b & a & d \\ 0 & 0 & 1 \end{bmatrix} \tag{33}$$

However, this solution should not be used in estimation of isometry, because the non-linear restriction in (31) is not enforced, i.e. $\exists\theta$ s.t. $\cos\theta = a$ and $\sin\theta = b$ is not guaranteed. From the SVD point of view a, b, c and d are completely independent variables. However, a similar formulation can be constructed for the similarity which actually yields to the form in (31). The difference is that $\exists\theta, s$ s.t. $s\cos\theta = a$ and $s\sin\theta = b$ is solvable $\forall a, b \in \mathbb{R}$. The solution is evident by summing the squared constraint equations:

$$s^2(\cos^2\theta + \sin^2\theta) = a^2 + b^2 \ , \tag{34}$$

from which $s = \sqrt{a^2 + b^2}$ taking the positive scale and θ obviously follows. The provided solution seems to produce equivalent results to Umeyama's method [18] in 2D. This restricted formulation in (32) is denoted as **DLT-R1** in the experiments. For comparison the experiments were also performed with the Umeyama's method (**UME**) [18].

A similar restricted formulation for affinity can also be defined:

$$\boldsymbol{H}_{\mathrm{A}} \Leftrightarrow \begin{pmatrix} \boldsymbol{h}^1 \\ \boldsymbol{h}^2 \\ \boldsymbol{h}^3 \end{pmatrix} = \begin{pmatrix} h_1 \\ h_2 \\ h_3 \\ h_4 \\ h_5 \\ h_6 \\ h_7 \\ h_8 \\ h_9 \end{pmatrix} = \begin{pmatrix} a_{11} \\ a_{12} \\ t_x \\ a_{21} \\ a_{22} \\ t_y \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \\ 0 \\ 0 \\ 1 \end{pmatrix} \quad , \tag{35}$$

which allows an arbitrary matrix $\boldsymbol{A}$ as in (1). The corresponding linear system is now

$$\begin{bmatrix} 0 & 0 & 0 & -w_i' x_i & -w_i' y_i & -w_i' w_i \\ w_i' x_i & w_i' y_i & w_i' w_i & 0 & 0 & 0 \end{bmatrix} \begin{pmatrix} a \\ b \\ c \\ d \\ e \\ f \end{pmatrix} = \begin{pmatrix} -y_i' w_i \\ x_i' w_i \end{pmatrix} \tag{36}$$

and from its solution the homography matrix is constructed as

$$\boldsymbol{H}_{\mathrm{A}} = \begin{bmatrix} a & b & c \\ d & e & f \\ 0 & 0 & 1 \end{bmatrix} \quad , \tag{37}$$

which obeys affinity for non-degenerate correspondences and is denoted as **LM-R2** in the experiments.

As already mentioned another restricted form used in the literature postulates a fixed element $h_9 = 1$ [1, 2, 9], but this should be avoided since for a projectivity the $h_9 = 0$ may appear in real situations and hence the homography matrix will be badly scaled and likely arbitrary. Also, for normalised data no robustness improvements can be gained from fixing an element [8]. Due to its historical significance the method is however included in the experiments (**DLT-R3**).

3.3.2 Data normalisation

The DLT method should not be used without the data normalisation [8]. If the different dimensions of data, including the homogeneous coordinate, which is typically set to unity, differ by orders of magnitude, an estimate minimising the algebraic error will be biased toward the largest dimension. Therefore, it is necessary to perform data normalisation before the SVD step and afterwards the estimated homography matrix must be denormalised. The standard normalisation method (isotropic normalisation) is described in [8]. Hartley and Zisserman also discuss the case, where the order of magnitude differences between data dimensions are desired, and present an alternative normalisation method. A DLT method utilising normalised data is prefixed with **NDLT**.

3.4 Other estimation methods

The literature contains several other linear or near linear estimation methods, such as Umeyama's [18] for similarity and isometry (scaling term is discarded) in multiple dimensions, but these seem to be for special cases or equivalent formulations to 2-D DLT (such as the Umeyama to DLT-R1). The most important other methods are clearly the iterative methods optimising other than the algebraic error.

Within existing non-linear and iterative methods the most popular approach is the gradient descent where the two important considerations are 1) which target (error) function should be optimised and 2) what gradient step method should be used in the optimisation. These questions are well covered in the book by Hartley and Zisserman [8], where the two important error functions are geometric distance and re-projection error. Since the iterative estimation however is not feasible in our case, these methods were not included to our study.

4 Experiments

As already mentioned several times, our main interests are methods for feature-based object detection, localisation, registration and recognition in 2-D. These methods utilise local object (category) specific features (for example, eye centres and nostrils of human faces) which are in the "constellation search" stage used as corresponding points (see, e.g., [3, 13]). Geometric transformations are needed to transform the detected image features to an object model space or vice versa. In this problem domain, the homography estimation is very challenging since objects are typically represented by a few local features whose spatial variation can be significant (e.g. natural spatial variation of facial landmarks between individuals). In practice, a good estimate should be achievable with near the minimum required number of correspondences or at least, as a rule of thumb, with less than 10 point correspondences. Moreover, for iterative search there are no computational resources available until at least a rough object pose estimate has been achieved (initial and refining stages in [13]).

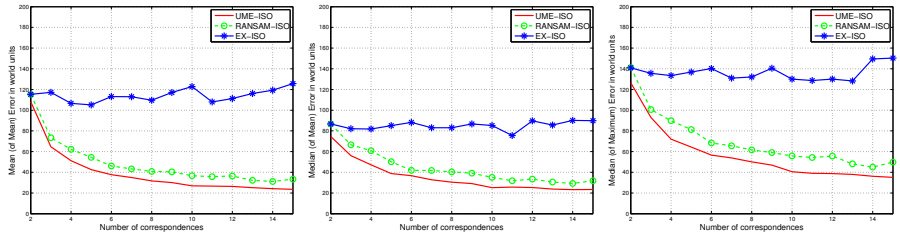
4.1 Simulated data

We report experimental results as a statistical estimation error for 200 times repeated tests of 100 randomly generated points. The measured estimation error was the geometric distance between the true target points and the true source points transformed using estimated homography, i.e. the true (forward) transfer error. We report mean of mean errors, median of mean errors and median of maximum errors. Mean over 200 repetitions of mean error of 100 points provides a general accuracy measure. The median is tolerant to a few ($< 50\%$) disastrous estimation failures, and therefore, may reveal methods which are more accurate but less robust. Similarly, the median of maximums provides the error upper bound for a successful estimation. The transformation parameters and the noise model were selected to simulate image capturing in practise. Points were evenly distributed to the span of the input space, 1 to 10 points randomly selected for estimation and the rest used for the error measure.

4.1.1 Isometry

In the first experiment, the world consisted of random points, $\boldsymbol{x} = (x, y)^{\mathrm{T}}$, drawn uniformly from $x \in [0, 1000]$ and $y \in [0, 500]$ (minimum length of all dimensions $l_{\min} = 500$). For each repeat, an orientation preserving isometry transformation as in (2) was generated from a uniformly random angle $-180^\circ \leq \theta \leq 180^\circ$ and uniformly distributed random translation $\|\boldsymbol{t}\| \leq 0.1 l_{\min}$. Point correspondences were randomly selected and the points were perturbed by white noise from $\mathcal{N}(0, (\gamma l_{\min})^2)$, that is, noise was defined by the standard deviation proportional to the dimensions of the source image (absolute error in pixels independent of resolution). The noise was separately added to the source (model) and target (image) points.

The results for Umeyama's method without the scale (UME-ISO), random sampled exact (RANSAM-ISO) and exact (EX-ISO) methods are shown in Fig. 1. As expected the exact method is comparable to others only with the smallest number of correspondences available ($= 2$). The random sampling exact method performed well even for a larger number of points as it selects the best combination according to the transfer error, but the Umeyama's method with the scale discarded outperformed other methods as it utilises all points in the estimation. All three methods were very robust as the mean and median graphs correspond to each other; no disastrous estimation errors occurred for the used noise level. The accuracy is further illustrated in Fig. 2.



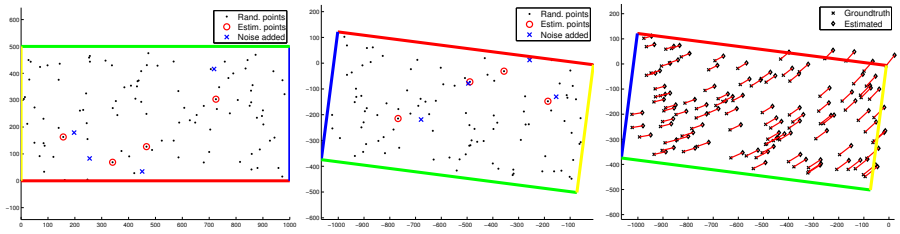
(a) (b) (c)

Figure 1: Isometry estimation accuracy (noise $\gamma = 0.080$): (a) average mean error; (b) median mean error; (c) median maximum error.

4.1.2 Similarity

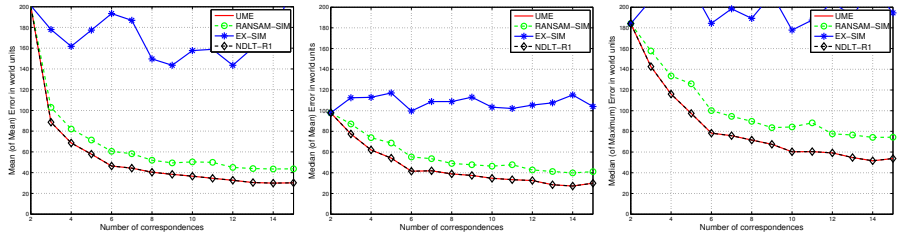
In this experiment, world data was generated exactly as in the previous, but in addition, the isotropic scale factor was randomly drawn from uniform $s \in \left[1/\sqrt{2}, \sqrt{2}\right]$ (at most half octave scaling).

The results for the similarity transformation estimation are shown in Fig. 3. In this experiment, Umeyama's method (UME) and normalised restricted DLT (NDLT-R1) provided practically the same estimates and were the best estimation methods. With the used error level the exact method (EX-SIM) provided poor accuracy, but still its random sampled version (RANSAM-SIM) was very good.



(a) (b) (c)

Figure 2: Example of isometry estimation with noise in corresponding points (UME-ISO): (a) 100 random points, 4 randomly selected estimation points and the same points with noise ($\gamma = 0.080$); (b) random transformation applied and the corresponding points with random noise added ($\gamma = 0.080$); (c) the point cloud after correct and estimated transformation (mean error 48.3 and maximum error 63.3 units).



(a) (b) (c)

Figure 3: Accuracy for similarity estimation (noise $\gamma = 0.080$): (a) average mean error; (b) median mean error; (c) median maximum error. UME and NDLT-R1 curves completely overlap.

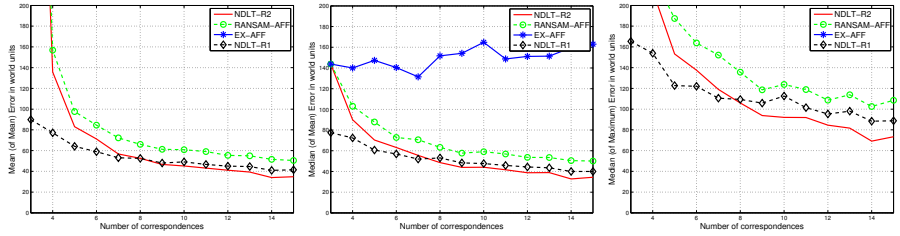
4.1.3 Affinity

In this experiment, the isotropic scale was replaced by two independent scale factors λ_1 and $\lambda_2 \in \left[1/\sqrt{2}, \sqrt{2}\right]$ $-180° \leq \theta \leq 180°$.

The experimental results for the error level $\gamma = 0.080$ are shown in Fig. 4. With the used noise level, the exact method became completely unfeasible providing only poor estimates. The normalised DLT restricted to affinity (NDLT-R2) performed the best, but another important result became evident: the more restricted homography estimation (similarity, NDLT-R1) outperformed all affinity estimation methods for a small number of correspondences (≤ 8) and the average performance remained good even further. Fig. 4 provides the first evidence how under-parametrisation can be beneficial for noisy data with few correspondences. The result is demonstrated in Fig. 5.

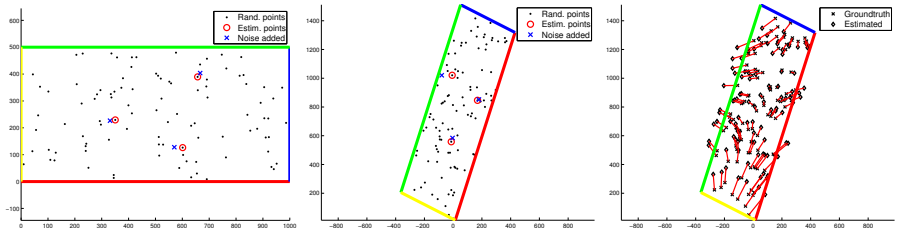
4.1.4 Projectivity

The parameters for the projectivity experiment were the same as for the affinity experiment, but in addition, the last row $\boldsymbol{h}_3$ of a homography matrix, the "elation transform



(a) (b) (c)

Figure 4: Accuracy for affinity estimation (noise $\gamma = 0.080$): (a) average mean error; (b) median mean error; (c) median maximum error.



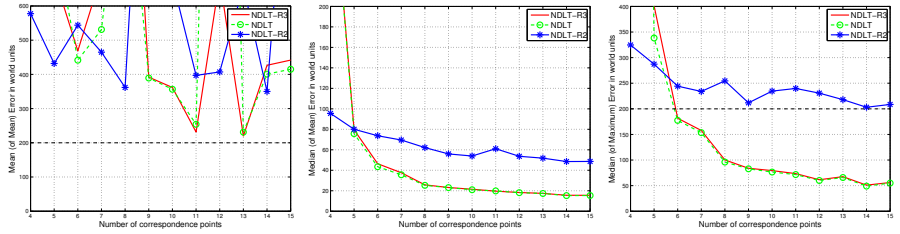
(a) (b) (c)

Figure 5: Example of affinity estimation with noise in corresponding points and similarity estimation method (NDLT-R1): (a) 100 random points, 3 randomly selected estimation points and the same points with noise ($\gamma = 0.080$); (b) random transformation applied and the corresponding points with random noise added ($\gamma = 0.080$); (c) the point cloud after correct and estimated transformations (mean error 75.1 and maximum error 181.2 units).

parameters" v_1, v_2 and v were randomly drawn from $v_1, v_2 \in [-0.001, 0.001]$ and $v \in [0.999, 1.001]$ (note that $v \neq 0$ is the scale of the matrix and should not affect to the estimated transformation). The elation parameters were allowed to vary in these limited intervals since the noise model was tied to the source space, and therefore, the target space dimensions should at least roughly correspond to the source dimensions. However, this corresponds to the typical situation in practise. The noise level was reduced to $\gamma = 0.025$.

The results of the projectivity experiment, shown in Fig. 6, were completely different as compared to any of the previous. Disastrous estimation errors occurred for all estimation methods, and therefore, the mean error graphs in Fig. 6(a) are completely out of limits. On the other hand, the median results, shown in Fig. 6(b), provided means to compare the different methods. Again, the over-restricted homography (affinity, NDLT-R2) outperformed other methods for 4 and 5 correspondences, its maximum errors, however, being severe. The normalised DLT methods NDLT and NDLT-R3 performed equally and since the NDLT also allows $v = h_9 = 0$ it should be preferred.

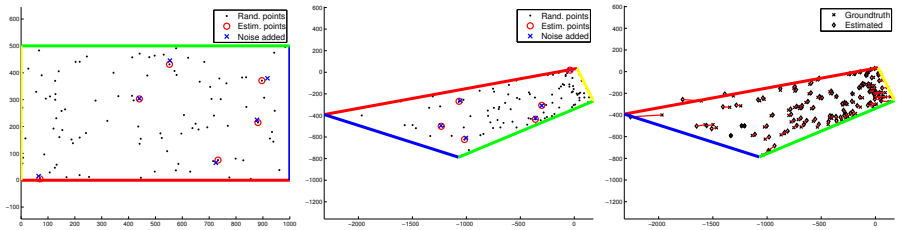
The accuracy of projectivity estimation using the linear methods is further illustrated in Fig. 7The reason for severe, even disastrous, estimation failures is evident from the



(a) (b) (c)

Figure 6: Accuracy for projectivity estimation (noise $\gamma = 0.025$): (a) average mean error; (b) median mean error; (c) median maximum error.

graphs and also in the illustration. If there are too few correspondences for projectivity, the estimate becomes "over-fitted" to the selected points. Over-fitting provides poor generalisation (extrapolation) capability, i.e. the geometric error grows rapidly outside the convex hull of the correspondences.



(a) (b) (c)

Figure 7: Example of projectivity estimation with noise in corresponding points (NDLT-R3): (a) 100 random points, 6 randomly selected corresponding points and the same points with noise ($\gamma = 0.025$); (b) random transformation applied and the corresponding points with random noise added ($\gamma = 0.025$); (c) the point cloud after correct and estimated transformations (mean error 32.1 and maximum error 291.3 units).

4.2 Real data: face detection and localisation

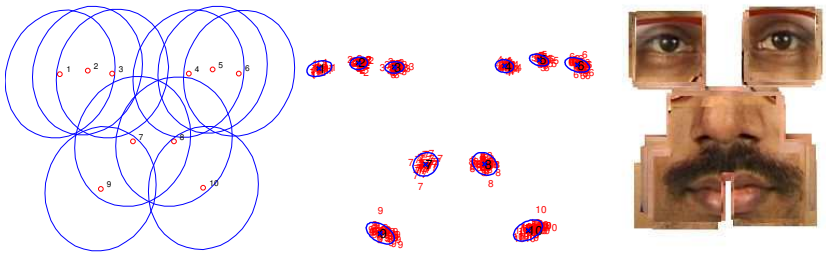
To verify that the previously presented simulated results correspond to practise we run similar experiments with 2 publicly available data sets of real objects (faces): XM2VTS/frontal (600 training and 560 test images) [15] and XM2VTS/non-frontal (XM2VTS MPEG7, 592 training and 588 test images). Links to the data and their ground truth (manually annotated 10 facial landmarks) are available at the project web page `http://www.it.lut.fi/project/homogr`. Example images from the both are shown in Fig. 8.

A training set is used to establish the spatial distributions of the landmarks, each of which is approximated by a Gaussian [13]. In image space the distributions are wide

(a) (b) (c) (d)

Figure 8: Example images from XM2VTS/frontal ((a) 341_1_2, (b)342_1_2) and XM2VTS/non-frontal ((c) 007__1_up, (d) 064_1_left) and 10 landmarks.

(Fig. 9 (a)), but in a proper model space they are well localised. An example of the set of model space distributions is shown in Fig. 9 (b), where a similarity transformation is used to transfer the landmarks into the model space. Any other homography class could be used instead of the similarity. To measure the fitness of a candidate landmark constellation in image space, it has to be transformed into the model space. This requires the estimation of a homography between the model and the landmarks in the image space.

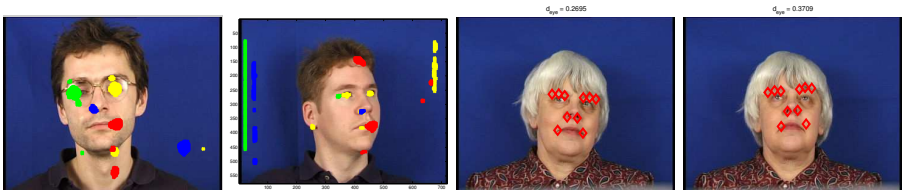


(a) (b) (c)

Figure 9: Point pattern mean model: (a) Mean and $2\times$ std deviation in the image space; (b) Mean and $2\times$ std deviation in the similarity normalised mean model space; (c) Local patches of XM2VTS/frontal transformed to the mean model space. [13]

The 10 landmarks were learned from a training set using local image feature extraction method proposed in [11]. For testing, the same method was used to extract the 100 best candidates for the each landmark from a test set (see Fig. 10 for examples). To facilitate comparisons the extracted landmarks have been made publicly available at `http://www.it.lut.fi/project/evex/`.

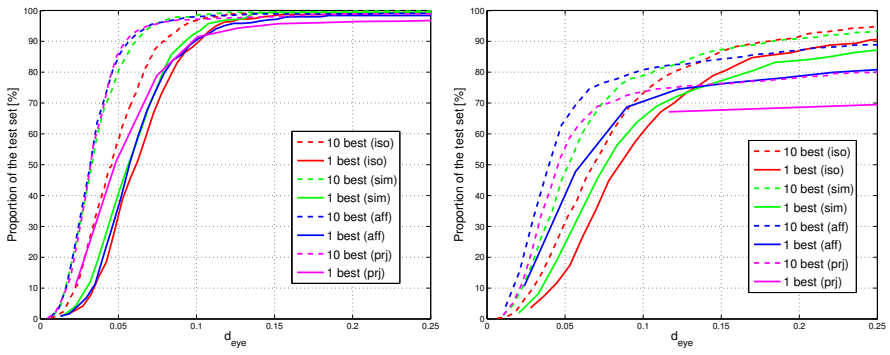
A simple but efficient point pattern based random sampling method for detecting the most likely hypotheses of an object presence was proposed in [13]. The method randomly picks a minimum number of correspondences from the set of extracted image features, estimates a homography to the mean model space and then transforms all extracted points to the mean model space. Then the best candidates for each landmark are selected and a statistical score formed by the product rule. We replaced the homography estimation part with the best methods according to the simulated experiment. The results for the both real data sets are shown in Fig. 11. The accuracy measure is the



(a) (b) (c) (d)

Figure 10: Examples of extracted image features using method in [11]: (a) XM2VTS/frontal (green: (image-wise) left outer eye corner, yellow: right eye centre, blue: left nostril, red: right mouth corner); (b) XM2VTS/non-frontal (c) XM2VTS/frontal failed face localisation (iso); (d) XM2VTS/frontal failed face localisation (iso).

d_{eye} distance which measures how far the eye centres are from the ground truth centres normalised by the eye distance (image resolution invariance) [12]. $d_{\text{eye}} > 0.25$ is considered as a detection failure. For the XM2VTS/frontal all methods performed very well, only isometry being notably less accurate using 10 best hypotheses of the object occurrence. This is explained by the fact that isometry cannot capture the scale variation of faces. For XM2VTS/non-frontal (Fig. 11(b)) affinity outperformed all others as the face rotation in-depth is best explained by affinity (face considered as a plane) and cannot be captured by the more restricted transformations. Similarity, however, was very good as well and projectivity failed as not enough correspondences were correctly detected in every image due to natural occlusion.



(a) (b)

Figure 11: Face localisation results (cumulative graphs for different homographies denoted by the different colours): (a) XM2VTS/frontal; (b) XM2VTS/non-frontal.

5 Conclusions

In this study we experimented on which fast (linear) methods should be preferred for homography estimation in 2D-to-2D object detection, localisation, registration and

recognition methods that are based on point correspondences. The results are especially important in cases where there are only a few correspondences and their locations in a model or the image space or both are significantly uncertain. That is the case in the most part-based object models. In addition, we demonstrated the fact that more restricted homographies provide increased robustness. As a rule of thumb, 2D-to-2D methods should go beyond affinity and utilise projectivity only in recognised special cases. Matlab implementations of all the used homography estimation methods are available at `http://www.it.lut.fi/project/homogr`.

As demonstrated using the simulated data the exact methods (EX-*) should be used only for the minimum number of correspondences. Otherwise UME-ISO should be used for isometry estimation, UME (NDLT-R1) for similarity, NDLT-R2 for affinity and NDLT for projectivity. However, if only a few correspondences are available and the correspondences are noisy, then the usage of the more restricted transforms provides more robust results. This fact was further tested by applying the selected methods to real experiments where only a negligible accuracy difference occurred between the homographies. However, for example for XM2VTS/frontal projectivity seems the most accurate in Fig. 11(a), but gave a failure rate of 1.07% as isometry only 0.36% (two images shown in Fig. 11(c)-(d)), i.e. projectivity generally provided a more accurate localisation, but isometry was more robust. It is noteworthy that using the more optimal linear methods for homography estimation from a few point correspondences we achieved significantly better results to [13] using otherwise the same methods.

References

[1] Y.I. Abdel-Aziz and H.M. Karara. Direct linear transformation into object space coordinates in close-range photogrammetry. In *Proc. Proc. Symp. Close-Range Photogrammetry*, pages 1–18, 1971.

[2] D.H. Ballard and C.M Brown. *Computer Vision*. Prentice Hall, 1982. Available on-line.

[3] A. Drobchenko, J. Ilonen, J.-K. Kamarainen, A. Sadovnikov, H. Kälviäinen, and M. Hamouz. Object class detection using local image features and point pattern matching constellation search. In *Proc. of the Scandinavian Conf. on Image Analysis (SCIA2007)*, pages 273–282, Aalborg, Denmark, 2007.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[6] M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Graphics and Image Processing*, 24(6), 1981.

[7] M. Hamouz, J. Kittler, J.-K. Kamarainen, and H. Kälviäinen. Hypotheses-driven affine invariant localization of faces in verification systems. In *Proceedings of the*

International Conference on Audio- and Video-Based Biometric Person Authentication, pages 276–284, 2003.

[8] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in computer vision*. Cambridge press, 2003.

[9] Janne Heikkilä. Geometric camera calibration using circular control points. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10), 2000.

[10] S. Helmer and D.G. Lowe. Object recognition with many local features. In *Workshop on Generative Model Based Vision*, 2004.

[11] J. Ilonen, J.-K. Kamarainen, P. Paalanen, M. Hamouz, J. Kittler, and H. Kälviäinen. Image feature localization by multiple hypothesis testing of Gabor features. *IEEE Transactions on Image Processing*, 17(3):311–325, 2008.

[12] O. Jesorsky, K.J. Kirchberg, and R.W. Frischholz. Robust face detection using the hausdorff distance. In *Proc. of 3rd Int. Conf. on Audio- and Video-based Biometric Person Authentication*, pages 90–95, 2001.

[13] J.-K. Kamarainen, M. Hamouz, J. Kittler, P. Paalanen, J. Ilonen, and A. Drobchenko. Object localisation using generative probability model for spatial constellation and local image features. In *Proc. of the ICCV 2007 Workshop on Non-Rigid Registration and Tracking Through Learning (NRTL2007)*, Rio de Janeiro, Brazil, 2007.

[14] David G. Lowe. Object recognition from local scale-invariant features. In *Proc. of the International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.

[15] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS Database. In R. Chellapa, editor, *Proc. of Second Int. Conf. on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[16] S. Savarese and L. Fei-Fei. 3d generic object categorization, localization and pose estimation. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, 2007.

[17] E. Trucco and A. Verri. *Introductory Techniques for Computer Vision*. Prentice Hall, 1998.

[18] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991.