An interactive method for the image alignment problem based on partially supervised correspondence

Xavier Cortés, Francesc Serratosa

Universitat Rovira i Virgili, Departament d’Enginyeria Informàtica i Matemàtiques, Spain

ABSTRACT

In a robotic scenario in which robots are localised through the images they obtain, it is usual that the current image registration techniques fail due to large differences between images, occlusions, shadows and so on. In these cases a human can interact on the computer vision system to help it to find a preliminary mapping between some salient points. We present a method to solve the alignment problem between two images in which a human (or oracle) imposes a partial correspondence between salient points extracted from the images through a Human–Machine Interface and the Interactive Correspondence Method computes a new image alignment and a set of point correspondences. Practical evaluation shows that in two human interactions our method finds an almost optimal alignment and correspondence set between images that any current completely automatic method is able to obtain. Moreover, we demonstrate that in situations that there is a large difference between images, adding interactivity is the only method that achieves an acceptable result.

1. Introduction

In recent years, interaction between robots and humans has increased rapidly. Applications of this field are very diverse, ranging from developing automatic exploration sites (Trevai et al., 2003) to using robot formations to transport and evacuate people in emergency situations (Casper & Murphy, 2003). Within the area of social and cooperative robots, interactions between a group of people and a set of accompanying robots have become a primary point of interest (Garrell & Sanfeliu, 2012).

One of the low level tasks that these systems have to face is the automatic recognition of scenes and objects the robot visualises. This problem is usually called image registration in the computer vision research field. Image registration tries to determine which parts of one image correspond to which parts of another image. This problem often arises at the early stages of many computer vision applications such as scene reconstruction, object recognition and tracking, pose recovery and image retrieval. Therefore, it has been of basic importance to develop effective methods that are both robust in the sense of being able to deal with noisy measurements and in the sense of having a wide field of application. Two interesting image registration surveys are (Zitová & Flusser, 2003; Salvi, Matabosch, Fofi, & Forest, 2007).

We present an interactive method such that a human can help the object or scene recognition modules incorporated on the robots when these modules consider they are not able to automatically solve the registration problem. This type of interaction is completely different from the ones presented in Trevai et al. (2003), Casper and Murphy (2003) and Garrell and Sanfeliu (2012) since in those cases, the human interaction is performed in a higher level. For instance, in Trevai et al. (2003), Casper and Murphy (2003), the interaction is based on imposing orders such as “move straight ahead” or “go up stairs”. In Garrell and Sanfeliu (2012), the orders are “follow this person” or “bring me to the exit”. Nevertheless, our experience has shown us that in most of the cases, robots cannot perform these orders due to they cannot solve the low-level registration problem. Therefore, interacting in this low-level task, that it is really easy for humans, frequently makes not necessary to interact on higher level tasks in which the interaction is more complicated do to the need of having more knowledge of the current situation such as position of other robots, automatically built map of the environment or current position of other objects. We have not found any paper that the aim is to interact on the registration problem integrated in a robotics’ framework. For this reason, we first explain the general robotics’ framework and then we explain the interactive registration method.
The three typical steps involved in the solution of the image registration problem are the following (Xu & Petrou, 2011). First, some salient points are selected from both images. Second, a set of tentative matches between these sets of points is computed together with the image alignment. And third, a process of outlier rejection that eliminates the spurious correspondences can further refine these tentative matches and the initial alignment.

Salient points, which play the role of parts of the image to be matched, are image locations that can be robustly detected among different instances of the same scene with varying imaging conditions. These points can be corners (intersection of two edges) (Tomasi & Kanade, 1991), maximum curvature points (Han & Brady, 2005) or isolated points of maximum or minimum local intensity (Rosten & Drummond, 2006). There is an evaluation of the most competent approaches in Mikolajczyk and Schmid (2005). When salient points have been detected, several correspondence methods can be applied that obtain the alignment (or homography) that maps one image into the other (Zhang, 1994), discards outlier points (Fischler & Bolles, 1981) or characterises the image into an attributed graph (Sanromà, Alquézar, & Serratosa, 2012; Sanromà, Alquézar, Serratosa, & Herrera, 2012).

The main drawback of image registration methods is that their ability to obtain the correspondence parameters strongly depends on the reliability of the initial tentative correspondences. Moreover, it is needed to jointly estimate the image alignment parameters and correspondence parameters. Considering the alignment parameters, there are two basic strategies. The first one is to consider a rigid deformation and the second one is to consider non-rigid deformation. In the first case, it is assumed the whole image (and so, the extracted salient points) suffers from the same deformation and so the image alignment parameters are applied equally to all the salient points or image pixels. Some examples are Luo and Hancock (2003), Rangarajan, Chui, and Bookstein (1997), Sanromà, Alquézar, Serratosa, and Herrera (2012) and Gold and Rangarajan (1996). In the second case, each salient points suffers a different projection and there are different alignment parameters applied to each salient point or image region. Some examples are Chui and Rangarajan (2003), Myronenko and Song (2010), Sanromà et al. (2012). Usually, the rigid strategy is applied to detect objects on outdoor images in which the deformation is mostly due to the change of the point of view. The non-rigid strategy is mostly applied to object detection or matching in medical or industrial images due to it is assumed objects suffer from deformations although the point of view is the same.

Humans are very good at finding the correspondences between local parts of an image regardless of the intrinsic or extrinsic characteristics of the point of view. Human interactivity on image registration has been applied on medical images (Khader & Ben Hamza, 2012; Pfluger, Thomas, et al., 2000; Pietrzyk et al., 1994) and two systems have been patented (Berg~Von, 2010; Gering, 2010). These papers and patents are really specific on some medical environments and for this reason cannot be applied on our problem. In Pfluger et al. (2000), they show a comparison of 3-D images on MRI-SPECT format and they concretise on images from the brain. In Pietrzyk et al., 1994), authors present a method to validate the 3D position given 3D medical images. Finally, in Khader and Ben Hamza (2012), the aim is to solve the registration problem given similar medical images extracted from different sensors or technologies. Patent (Berg~Von, 2010) defines a system for registration thorax X-ray images such that it does not depend on bony structures. And patent (Gering, 2010) defines a multi-scale registration for medical images where images are first aligned at a coarse resolution, and subsequently at progressively finer resolutions; user input is applied at the current scale.

Current automatic methods to extract parts of images and their correspondences in non-controlled environments are far away of having the performance of a human. Fig. 1 shows two images extracted from RESID database (http://www.feature-space.org). In each image 50 salient points have been extracted by method (Harris & Stephens, 1988). Outlier detector (Fischler & Bolles, 1981) has considered 43 salient points where outliers and only 7 where inliers. The correspondence detector has missed 6 of the 7 points (red lines) and only has hit 1 point (green line). This is because of the large differences between both images and more precisely, due to the lack of ability of the initial correspondence detector to find a good initial correspondence.

For this reason, in this paper, we propose a semi-automatic method in which humans can interact into the system when it is considered the quality of the automatically found correspondences is not good enough and then they impose a partial and initial correspondence between some local parts of two scenes. To do so, the oracle selects through a Human–Machine Interface the points of both images that have to be mapped.

The oracle basically helps the pattern recognition system to find the initial correspondence through its interaction. In this action, there is a learning process in which the pattern recognition system learns from the oracle some node correspondences (local knowledge). Note that this kind of learning is different from the one presented in Caetano, McAuley, Cheng, Le, and Smola (2009), in which the pattern recognition system learns some weights (global knowledge) to find the graph node correspondences more accurately. Therefore, our framework is the Partially Supervised Learning (PSL) since only few point correspondences are learned from the oracle. PSL is a research field of machine learning and pattern recognition that builds upon the broad areas of learning, clustering and estimating classifiers from partially (or, weakly) labelled data. PSL studies are particularly expected to involve the identification of scenarios for successful real-world applications. To state just a few examples, significant instances of PSL applications are rooted in image and multimodal information processing, sensor and information fusion, human machine interaction and data mining. Recently a special issue has been published about PSL for Pattern Recognition in which several methods and applications are described (Schwenker, in press-a,b).

Human–Machine Interface (HMI), also named Human–Computer Interface, provides more natural, powerful, and compelling interactive experiences. For decades, HMI has been an active research field closely related to new technology advances. Paper published in Matthew (2014) provides an interesting review of the current trends and key aspects of HMI, for instance in non-desktop computing. Some of the latest applications have been natural language understanding (Revuelta-Martinez, Rodriguez, Garcia-Varea, & Montero, 2013), semi-supervised database indexing (Phuong, Visani, Boucher, & Ogier, 2013), image segmentation (de Miranda, Falcão, & Udupa, 2010; Panagiotakis et al., 2013) or handwritten text transcription (Romero, Toselli, & Vidal, 2012) between others. Finally, human interactivity specifically applied to pattern recognition has been considered in Toselli, Vidal, and Casacuberta (2011).

Considering the visual presentation of the HMI, three main approaches have been proposed in the literature: (a) side-by-side views (Andrews, Wohlfahrt, & Wurzinger, 2009; Holten & Wijk, 2008) (b) superimposed or merged views (Ahberg, 1996; Erten, Harding, Kobourov, Wampler, & Yee, 2003) and (c) animations (Diehl & Gorg, 2002; Yee & Fisher, 2001). These approaches are often complemented with techniques for highlighting the correspondences between nodes of both graphs such as lines that go from one graph to the other (Collins & Carpendale, 2006; Holten & Wijk, 2008) or node colour coding (Elmqvist, Dragicievic, & Fekete, 2008; Kleinbe, 2002).

Fig. 2 shows, through the visual presentation of our HMI, the results of our Interactive Correspondence Method (ICM). We have...
chosen the side-by-side view and highlighting based on colour lines. In this case only two interactions where needed (yellow lines) to achieve 10 correct correspondences (green lines) and only one un-correct correspondence (red line). Note that not only the number of correct correspondences has increased but also the number of inliers has increased from 7 to 11.

We advocate a cooperative model in which robots have cameras and they have to identify objects and scenes in a cooperative manner with the aim of detecting specific objects or performing Simultaneous Localisation And Mapping (SLAM). We assume there is a master system that receives images from robots and finds the correspondences between images of these robots to perform higher-level tasks. These images may have been taken from different scenes or from the same scene but with completely different points of view, illuminations and so on. When scenes or objects from scenes have been localised, the master system sends new orders to the robots such as specific movements, grasping objects and so on. When a robot is not able to recognise the scene, then it stops and asks to the master system to localise it or send it the information of the scene. The master obtains the whole information through the correspondence between the current and past images of all the robots. Due to the environment (for instance, rescue inside buildings), we assume it is not possible to localise the robots through GPS, 3D triangulation or sensors on the wheels.

When we put into practice our system, we realised most of the cases that robots did not correctly react to the humans orders where due to they did not solve appropriately the low-level image registration problem. Therefore, the strength of our method is that putting the human interaction into the method, most of these non-correct robot reactions is solved. Nevertheless, technology tends to make systems to run as much autonomous as possible. For this reason, the main weakness of our method is that robots are less autonomous. Better registration methods are going to be discovered, and then, less need of human interaction is going to be needed.

A similar interaction method was presented in Villamizar, Andrade-Cetto, Sanfeliu, and Moreno-Noguer (2012). In that case, there is only one robot and the human decides if a selected part of the image is a face and in the case that it is, imposes the name of the person.
Fig. 4 shows a schematic view of our method based on an Interactive Correspondence Method and a Human–Machine Interface. The user visualises both images together with the salient points and the current correspondence between them. Then, the user imposes some point mappings and the output of the module is a correspondence in which the oracle has influenced. This process is repeated until the user considers the correspondence is good enough. A preliminary version of this method was presented in Cortés, Moreno, and Serratosa (2013). In that paper, authors presented a simple interactive method to estimate the homography between two images.

Besides, this project is part of a larger project in which social robots guide people through urban areas (Garrell & Sanfeliu, 2012) and they have tracking abilities (Serratosa, Alquézar, & Amézquita, 2012). Fig. 5 represents three robots performing guiding tasks in an indoor environment. Robots fence the visitor group to force them to follow a specific tour. Robots need to work in a cooperative manner to keep a triangular shape in which people have to be inside. In these cooperative tasks, it is crucial to have a low-level computer vision task such that the salient points extracted from the three robots are properly mapped. In this environment, there is a human that, through our HMI, helps to properly map the salient points.

The rest of the paper is organised as follows. In the next section, we succinctly explain the method used to deduct an alignment given a correspondence between two sets of points. In Section 3, we explain our method to achieve an automatic alignment and correspondence and in Section 4, we add interactivity to this method. This means that a human proposes some point mappings and the method finds the best image alignment and point set correspondences considering this human imposition. In Section 5, we experimentally validate our interactive model and show that our method solves the alignment problem on images with large differences. We conclude the paper in Section 6.
and a set of pairs corresponding between $d$ and $d^+$ and $(d, y) = \{x, y\}$. We consider the homography (points alignment) that transforms one set of points (represented as vector $x$) to another set of points (represented as the other $y$). Note that, in some applications, deformations are not rigid (Serratosa, Cortés, & Solé-Ribalta, 2013) and therefore, the method that we present cannot be applied.

Given two images $I_1, I_2$ and a set of pairs corresponding points on these images $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ it is possible to obtain a transformation matrix $F$ that projects each of the points in $I_1$ into the points in $I_2$ with a minimal error using the Direct Linear Transformation algorithm. Then $[x_i', y_i', 1]^T = F[x_i, y_i, 1]^T$, where $(x_i', y_i') \in I_2$ and $(x_i, y_i) \in I_1$. Throughout this paper we suppose the transformations images suffer to be aligned to the other ones are affine transformations.

If this is not the case, the only part that has to be reconsidered is the error function we arrive below and therefore the obtained matrix $A$ and vector $B$.

If we suppose that the mean error has a normal distribution, then the least-square estimation is optimal (Pedhazur & Elazar, 1982). Moreover, if we assume an affine transformation, then the transformation matrix is defined as

$$ F = \begin{bmatrix} a & -b & c \\ b & a & d \\ 0 & 0 & 1 \end{bmatrix} $$

where $a = S \cos(\alpha), b = S \sin(\alpha), S$ = scale, $c$ = translation in $x$ and $d$ = translation in $y$.

In this case the error function is usually estimated as follows:

$$ E(a, b, c, d) = \sum_{i=1}^{n} \left[ (ax_i - by_i + c - x_i')^2 + (bx_i + ay_i + d - y_i')^2 \right] $$

We find the minimum through deriving the function; $\frac{df}{da} = 0, \frac{df}{db} = 0, \frac{df}{dc} = 0$ and $\frac{df}{dd} = 0$. And we arrive to the following linear problem, $A(a, b, c, d)^T + B = 0$.

Where,

$$ A = \begin{bmatrix} \sum x_i^2 + y_i^2 & x_i & y_i & 0 \\ 0 & \sum x_i^2 + y_i^2 & -y_i & n \\ \sum y_i & \sum y_i & n & 0 \\ \sum x_i & \sum y_i & 0 & n \end{bmatrix}, \quad B = \begin{bmatrix} \sum x_i y_i' - y_i x_i' \\ \sum x_i y_i' - x_i y_i' \end{bmatrix} $$

We solve this linear system using LU factorization (Bunch & Hopcroft, 1974).

### 3. Automatic Alignment & Correspondence Estimation

Fig. 6 shows the main steps of our Automatic Alignment Estimation. Any interest point extraction algorithm like, Harris and Stephens (1988), DoG (Lowe, 2004) or LoG (Lindeberg, 1998; Mikolajczyk & Schmid, 2004) can be used to obtain the salient points $P_i$ and $P_j$, given images $I_1$ and $I_2$. While these salient points are obtained, the system obtains the homography $H_{1,2}$ between them (alignment between points), for instance, using the Iterative Closest Point algorithm (Zhang, 1992) or RANSAC algorithm (Fischler & Bolles, 1981). In cases that the algorithm not only obtains the homography but also detects the outliers, the algorithm returns a Consistent Set $CS_{1,2}$, for instance the RANSAC algorithm (Fischler & Bolles, 1981). This consistent set is composed by two sets of points (one from each initial salient point sets) that the algorithm considers they have to be mapped.

In our interactive framework (Section 4), we need to validate the returned alignment. For this reason, we propose to project the set of points $P_i$ and image $I_1$ to obtain $P_i'$ and $I_1'$ where $P_i' = H_{1,2} \cdot P_i$ and $I_1' = H_{1,2} \cdot I_1$. Thus, the human can visualise the output through a Human–Machine Display. Moreover, the display also shows the consensus set generated by the alignment method (green lines). Note that there is no human interaction in this framework, but we present the interactive framework in the next section.

Fig. 7 shows the basic steps of our Automatic Correspondence Estimation. In this case, three steps (red box) have been added to...
obtain the final correspondence \( f_{1,2} \). In the first one, a similarity matrix is defined with the projected points \( P_1 \) and the original points \( P_2 \). Then this matrix is modified with the information of the Consensus Set \( CS_{1,2} \) generated by the homography estimator. The method considers the similarity is zero if the mapped points are not in the Consensus Set. For this reason, the whole cells that relate non-mapped points are set to zero. This method is inspired in the algorithm presented in Serratosa, Cortés, and Solé-Ribalta (2012) although in that case, the interactive algorithm modified a cost matrix instead of a similarity matrix. Algorithm 1 shows the process used to update the similarity matrix encapsulated in the Consensus Set Filtering module.

Algorithm 1. Consensus Set Filtering

**Input:** Similarity Matrix \( M_{1,2} \), Consensus Set \( CS_{1,2} \)

**Output:** Similarity Matrix with Restrictions \( M'_{1,2} \)

For all \( p'_1 \notin CS_{1,2} \) and \( p'_1 \in P_1 \)

\[ M'_{1,2}(i, b) = 0; \ \forall b \text{ being } p'_1 \text{ the } i\text{th element in } CS_{1,2} \]

End for all

For all \( p_2 \notin CS_{1,2} \) and \( p_2 \in P_2 \)

\[ M'_{1,2}(a, j) = 0; \ \forall a \text{ being } p_2 \text{ the } j\text{th element in } CS_{1,2} \]

End for all

end Algorithm

The last step obtains the final correspondence using an algorithm that solves the assignment problem such has the Munkres’ algorithm (Kuhn, 1955). In this case, the Human-Matching Display not only shows the projected images and points but also the found points correspondences \( f_{1,2} \) (green lines). Therefore, in this case the green lines are not the Consensus Set returned by the alignment algorithm but the final output correspondences. Again, there is no human interaction.

4. Interactive Alignment & Correspondence Estimation

Fig. 8 presents the main scheme of the Interactive Alignment Estimation. It is based on the alignment estimation model shown in Fig. 6 but the oracle’s interaction is added (dashed lines in the scheme). The main idea is to project the initial set of points \( P_1 \) and \( P_2 \) into \( P'_{1,2} \) and \( P''_{1,2} \) such that the input of the alignment estimation becomes “easier” and so the output alignment tends to be more accurate or more in concordance with the user. As commented in the introduction, the initial concordance estimation is crucial to obtain a good alignment and so our proposal is to modify and improve this initial set.

Fig. 9 shows the main scheme of the module called Alignment Estimation given the human Correspondence. The oracle imposes some correspondences between points \( f_{1,2} \). Note that \( f_{1,2} \) is usually a partial labelling between points of few point mappings. It is normal that in each interaction, the user imposes only one point
mapping although \( f_{1,2} \) represents the union of the whole imposed mappings through all interactions. We suppose there are no contradictions between oracle’s feedbacks. The Direct Linear Transform algorithm (Section 2) obtains homography \( H_{1,2} \). The Points Set Filtering given Images obtains two reduced point sets \( P_{1} \) and \( P_{2} \) from point sets \( P_{1}^{I} \) and \( P_{2}^{I} \). The discarded points in \( P_{1} \) and \( P_{2} \) are the ones that have to be outliers in \( I_{1}^{I} \) and \( I_{2}^{I} \). When we project \( I_{1} \) and obtain \( I_{1}^{I} = H_{1,2} \cdot I_{1} \), then \( I_{1}^{I} \) has the same information than \( I_{1} \) but the image is shown in the coordinate space of \( I_{2} \). Thus, we consider that a point in \( P_{1}^{I} \) (or \( P_{2}^{I} \)) is an outlier if its coordinates are out of the domain of image \( I_{2} \) (or \( I_{1}^{I} \)).

Fig. 10 graphically shows an example of the Points Set Filtering given Images. The last image represents the large image in which we have marked the salient points and also we have put on it the small image. Thus, blue points are inliers since they are inside the projected small image. Contrarily, red points are outliers.

Fig. 11 shows the main scheme of our Interactive Correspondence Estimation model. It is based on the Alignment Estimation Model shown in Fig. 8 but in this case, the feedback of the oracle influences on two modules. The first one, called Alignment Estimation given a Correspondence, has just been commented (Fig. 9). The other one is the Feedback Set Filtering. This module executes algorithm 1 as it does the Consensus Set Filtering module but the inputs are the feedback of the human \( f_{1,2} \) and matrix \( M_{1,2} \).

We force a rigid transformation between images but non-rigid transformations appear in real applications. For this reason, if we make the oracles’ feedback to only influence on the Alignment Estimation given the human Correspondence and the Feedback Set was considered to be null in the Feedback Set Filtering module, the model may not achieve the point correspondences imposed by the oracle. That is, in cases that there is no a real rigid transformation, it may happen that \( f_{1,2} \neq f_{1,2}^{I} \) although the oracle has imposed the correspondence between all points. For this reason, it is needed the Feedback Set Filtering. We show this effect in the practical evaluation.

5. Practical evaluation

To validate our method, we have performed two different kinds of experiments. The aim of the former ones is to validate the Interactive Alignment Estimation whereas the aim of the latter ones is to validate the Interactive Correspondence Estimation. In the first case, it is needed to have an alignment ground truth (homography between images) but in the second case it is needed to have a correspondence ground truth (mapping between salient points of the images). For this reason, we have used two kinds of databases. In Section 5.1 we succinctly explain the used databases and their principal characteristics. In Section 5.2, we show the experiments applied on the Interactive Alignment Estimation and on the Interactive Correspondence Estimation.

5.1. Databases

We used four different databases to validate the Interactive Alignment Estimation model: BOAT, EAST PARK, EAST SOUTH and RESID and other four databases to validate the Interactive Correspondence Estimation model: APARTMENTS, HOUSE, HOTEL, FACES.

BOAT, EAST PARK, EAST SOUTH and RESID (http://www.featurespace.org): These databases are used to evaluate point extraction algorithms and image registration methods (Fig. 15)
shows some images from these databases). They consist of images taken from different outdoor scenarios across different zooms and rotations. Each database contains 10 or 11 images that are ordered according to a variation of zoom and rotation. Moreover, the homography matrix from the first image to the other ones is given. Points have been extracted using the scale-invariant feature detector (Lowe, 2004). For each image, we keep the 50 points with the highest scales. There is no correspondence ground truth.

APARTMENTS database (http://deim.urv.cat/~francesc.serratosa/databases/): It consists of ten images of an apartments building taken from different perspectives (Fig. 12.a). Twelve salient points have been manually extracted from each image. Each point...
represents the same part of the scene through the whole images, and so, the point correspondence is implicit in the point enumeration. The first 11 points represent parts of the building and the 12th point is the highest part of a lamppost.

HOTEL [http://www.featurespace.org], HOUSE [http://www.featurespace.org] and FACES (Oliver Jesorsky, Kirchberg, & Frischholz, 2001): Similarly to the APARTAMENTS database, they are composed of a set of images (Fig. 12b–d) and from each image, a set of points have been selected such as each point represents the same part of the scene through the whole images. Again, the correspondence between points is implicit in the point enumeration. Nevertheless, due to the degree of image distortion is very low, we generated other sets of points by artificially adding some noise on the original points (similar to Cortés and Serratosa (2013)). First, we added a global artificial transformation for all points (rotation and translation noise). And second, we added an individual noise for each point following a normal distribution. This noise simulates different viewpoints highly separated.

Fig. 10. Example of the Points Set Filtering given Images.

Fig. 11. Scheme of Interactive Correspondence Estimation model.
between them and makes the oracles interaction useful to find a good solution since the transformation between the original frame and the artificial one is non-rigid. Note that by construction, we do not guarantee the ground truth correspondence achieves the minimum cost.

5.2. Experiments

In all experiments we use Random Sample Consensus (Fischler & Bolles, 1981) as the Automatic Alignment Estimation algorithm.

5.2.1. Interactive Alignment Estimation

To evaluate the Interactive Alignment Estimation (IAE) we made two different experiments using the BOAT, EAST PARK, EAST SOUTH and RESID sequences. In the first one, we compared the IAE to the results presented in Sanromà et al. (2012a). In that paper, authors presented a comparative study of six completely automatic methods: Smooth Point Registration Using Neighbouring (Sanromà et al., 2012a), Softassign Procrustes (Rangarajan & Chui, 1997), Unified Approach (Luo & Hancock, 2003), Dual-Step Method (Cross & Hancock, 1998), Graduated Assignment (Gold & Rangarajan, 1996) and Matching by Correlation (Sanromà et al., 2012a).

![Fig. 12.](image1.png) One image extracted from (a) APARTMENTS, (b) HOUSE, (c) HOTEL, and (d) FACES databases.

![Fig. 13.](image2.png) (a) Mean End Point Error in pixels. (b) Number of obtained inliers. Interactive Correspondence Method with no interactions, Interactive Correspondence Method after two interactions, Smooth Point Registration Using Neighbouring (Sanromà et al., 2012a), Softassign Procrustes (Rangarajan & Chui, 1997), Unified Approach (Luo & Hancock, 2003), Dual-Step Method (Cross & Hancock, 1998), Graduated Assignment (Gold & Rangarajan, 1996) and Matching by Correlation (Sanromà et al., 2012a).

![Fig. 14.](image3.png) (a) Mean End Point Error in pixels and (b) number of correspondences respect of number of interactions. Boat, Eastsouth, Eastpark and Resid.
As mentioned by the authors, the image registration was performed only between adjacent images due to non-adjacent images presented non-admissible results. Concerning our method, we show the results when no user interactions are performed and at the second interaction. We suppose the oracle imposes node mappings that are in concordance with the alignment ground truth. That is, in each iteration, the oracle selects the point in the first image that has not been selected before and obtains the minimum Euclidean distance between its projection to the second image and the points in this second image.

Similarly to Sanromà et al. (2012a), we evaluate the goodness of the methods through the End Point Error, which is obtained as the sum of the Euclidean distances between points of the second image and points of the first image projected with the estimated alignment (homography). As good as the estimated alignment is, smaller have to be the distance between the points of the second image and projected points of the first image.

Similarly to Sanromà et al. (2012a), we evaluate the goodness of the methods through the End Point Error, which is obtained as the sum of the Euclidean distances between points of the second image and points of the first image projected with the estimated alignment (homography). As good as the estimated alignment is, smaller have to be the distance between the points of the second image and projected points of the first image.

The aim of the second experiments is to show the goodness of our method when there exist an important difference between the images to be considered. To do so, we have used the same datasets as the previous experiments but the whole combinations between images were considered (not only the adjacent ones). We do not present results of completely automatic methods since there were not acceptable. Fig. 14 shows the evolution of (a) the Mean End Point Error and (b) the number of inliers respect of the number of iterations. We can see that at the second iteration both measures tend to stabilize which means our method only demands to the user to interact two times. Moreover, while comparing the values of Figs. 13a and 14a we realise there is not a large difference between error values. Therefore, we conclude we achieve a good alignment although there is a large difference between the images to be considered. Fig. 15 shows the correspondences using our method but without interactions (images above) and after two mappings have been imposed (images below). Yellow lines: imposed mappings. Green lines: correct mappings. Red lines: incorrect mappings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 15. Some examples of matches from BOAT, EAST PARK, EAST SOUTH and RESID databases. Images above show the results without human interaction and images below show the results with two imposed interactions. Yellow lines: imposed mappings. Green lines: correct mappings. Red lines: incorrect mappings. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Fig. 14 shows a steep descend between iteration 1 and 2. This is because the interactive module establishes an acceptable alignment at iteration 2 that allows having a good initialization for RANSAC algorithm (Fischler & Bolles, 1981). It is very important for the RANSAC algorithm to have a good initialization because it can properly decide the outlier points. For this reason, there is an important increase in the number of found correspondences at the second interaction.

5.2.2. Interactive Correspondence Estimation

The aim of the experiments performed on the Automatic Correspondence Estimation is to minimize the Hamming distance between the obtained correspondence and the correspondence ground truth. We suppose the oracle imposes node mappings that are in concordance with the correspondence ground truth and, in each iteration, the oracle randomly selects one of the points that has not been selected before.

Fig. 16 shows the average Hamming distance respect the number of iterations in the four databases. In iteration 0 and 1, the Alignment Estimation given a Correspondence is not used due to two matchings are needed to deduce a unique affine alignment. Nevertheless, in iteration 1, the first node mapping imposition is introduced into the Feedback Set Filtering module. Interaction 2 is the first one that the whole model is applied. We have performed three different experiments. In the first ones (blue curve), the complete interactive method is used. In the second ones (orange curve), the Alignment Estimation given a Correspondence is not used and the human impositions only affect on the Feedback Set Filtering module. Conversely, in the third ones (yellow curve) the Alignment Estimation given a Correspondence is used but does not the Feedback Set Filtering module.

While using ( ), the model establishes an initial alignment at second interaction, for this reason the Hamming distance presents a steep descent. Throughout the other iterations, the model slightly updates the alignment and imposes the restrictions and so the Hamming distance decreases until arriving to zero. When Feedback Set Filtering module is deactivated ( ), the Hamming distance does not arrive to zero. This is because, in a real application, a rigid homography does not exist, which would allow to deduct the ideal correspondence applying only this transformation. Finally, in the case that Alignment Estimation given a Correspondence is deactivated ( ), the Hamming distances arrives at zero but more slowly than using the whole method. Through these experiments we show the need of using the whole method since one of the most important qualities that an interactive method needs to have is the ability to react in few oracle’s interactions.

6. Conclusions and future work

When a team of robots needs to have consistent shared information of its environment based on its computer vision systems, usually the completely automatic methods fail to find a good enough correspondence between the different parts of the scene. This is because of large differences of the seen images. In these cases, an interactive method can find a correspondence establishment (or improve a low quality one) and so increase the consistency of the shared information.

We have presented an interactive method that, given two images taken from different robots, the oracle imposes some correspondences between their local parts to estimate an initial alignment and to restrict the point correspondences. The main difference between our method and the methods presented in the literature related on robotics is that the interaction is performed at the low-level image registration process. In the other presented methods, the interaction is based on giving orders at high-level tasks.

Results show that in the first two iterations, there is an important increase on the quality of the correspondences. This result is important since it means it is not necessary a long-term dependence on the oracle but it is crucial on the initial stage. Therefore we conclude that with only two human interactions most of the situations that an automatic system cannot solve are now solved and obtaining good results. Our method has been developed in conjunction with a social robot network project (Garrell & Sanfeliu, 2012). Thanks to our iterative model, some new robot tasks have been achieved in which the robot images had low quality due to shadows, partial occlusions or extremely different point of views and illuminations. These good practical results validate...
our theory that it is useful the human interaction not only on high-level tasks but also on low-level tasks. Although there is a tendency to make the systems autonomous, as commented in the introduction section, some interactive methods and patents have been presented recently related on image registration. This is because, although this problem has been studied for a long time, in some applications or environments, their achievements are far away from being acceptable. Therefore, it is worth to introduce some level of human interaction, although the run time can be affected, to assure an acceptable accuracy. In a near future, it is supposed the interaction is going to be reduced when the automatic methods are going to increase their quality.

Nowadays, we are modelling the interaction on several images (or sets of points) simultaneously (not only one-to-one but many-to-many) with the aim of obtaining a coherent set of alignments and correspondences between different scenes. Moreover, we are also studying models for non-rigid homographies. Finally, we want to apply active query strategies as it was done in Cortés and Serratosa (2013) and Cortés, Serratosa, and Solé (2012) to suggest the points to be imposed by the oracle. We also want to reduce the computation run time by applying the Fast Bipartite method (Serratosa, 2014).

References

Cortés, X., & Serratosa, F. (2011). Active-learning query strategies applied to select a graph node given a graph labelling. In GI2013 (pp. 61–70). Austria, Vienna.
Diehl, S., & Gorg, C. (2002). Graphs, they are changing dynamic graph drawing for a near future, it is supposed the interaction is going to be reduced. In Proceedings of the 4th Alvey vision conference (pp. 147–151). Cambridge University Press.


