Pose Estimation Using Visual Entropy

Jianjun Gui*, Dongbing Gu and Huosheng Hu
School of Computer Science and Electronic Engineering, University of Essex
Wivenhoe Park, Colchester CO4 3SQ, UK. Emails: jgui@essex.ac.uk

Abstract—The trend of using visual information for pose estimation of camera becomes increasingly popular and diverse. In this paper, we propose a pose estimation based on visual information entropy. The constructed cost functions are robustness by natural and suitable to be applied in pose estimation in highly agile platforms. Especially, using the visual entropy or salient features as observation causes dramatically difference in computational time. The experiments based on real data from office environment shows that the entropy-based pose estimation using mutual information has huge potential in performance.

Index Terms—Pose tracking, Entropy, Mutual information

I. INTRODUCTION

As an easy means to observe the world, visual sensing has been widely adopted in all kinds of platform. Nowadays, it is very normal to find a tiny but high quality camera in our smart phones, laptops or even some electronic toys like drones. For academic research, “What can we get from the images?” is always ranked as a hot topic and has been developed to various branches according to different applications scenarios.

In computer vision community, camera poses estimation are often completed by the bundle adjustment [1] which can refine the environmental structure and camera poses at the same time. Visual odometry (VO) estimates 3D motion of camera sequentially, while in simultaneous localisation and mapping (SLAM), a global consistent map also should be built [2].

Among the trend of visual based pose estimation, there exists two major methods to deal with visual information. The most standard one is to extract a sparse set of salient features, then express them using feature descriptors. Although the feature-based methods come with an obvious limitation, i.e. only the scene contains a certain feature pattern (e.g. corners, lines) can be expressed and used, the successful rotation and scale invariant characters of some feature detectors and descriptors still make methods in visual SLAM or visual odometry (e.g. PTAM [3], monoSLAM [4]) be commonly accepted.

On the other hand, the direct methods recover the structure or motion directly from the intensity and/or gradient of the sequential images, on the occasion, the magnitude and/or direction of pixels can be used in pose estimation compared to only sparse features. The most prominent character of direct based methods is that all information in one image can be exploited, even for the environments with little texture, few keypoints or huge impact of camera defocus and blur. Only recent years, some direct based visual odometry methods have been proposed and become popular. Methods in [5], [6] built a fully dense depth maps on per pixel basis and the computational ability of state-of-art GPU were adopted. Researchers in [7] cut down dense region to reduce the computational burden and proposed a semi-dense depth mapping method. The method in [8] combined direct information with keypoints repetitively to the enhance intra-frame tracking results.

As the value of intensity is quite sensitive to illumination change, direct based methods would trigger large difference in pixels even with small movements, which can lead to divergence in algorithm. However, the statistic information reflected by the intensity is more stable than the intensity value itself. And it happens that the conception of entropy in information theory gives the researchers a clue to deal with the intensity information. Considering two images as different signal sources, the alignment between them can be expressed by the value of mutual information. This idea has been widely used in image registration [9] or visual servo [10].

In this paper, we present two different entropy-based expressions for pose estimation. While one adopts direct information on per pixel basis, the other uses the traditional features as measurements. Experimental results show the efficiency and robustness for pose estimation and further mapping or global alignment. It is also revealed that the method using mutual information on the whole image is superior in computational complexity when compared to feature based methods.

This paper is structured as follows: in the first section, a generalized projection of spatial point is abstracted as a basic model, followed by an overview of the entropy regarding visual information as signals in section III. In section IV, a cost function is constructed using the entropy. Some experiments based on real data are analysed in the aspects of robustness, stability and efficiency in section V.

II. A GENERAL PROJECTION MODEL

A camera motion in free space can be modelled as in Figure 1. The two images are sequential where the motion $\xi$ is small and rotation is remitted, or large span in space
where $\xi$ contains rotation and translation. A point $u$ in the image is regarded as a pixel or a extracted feature and its paired $u'$ is in the second image. Both of the points together are coplanar, forming geometry constraints for pose estimation. In order to reduce the computational complexity, the transformation is usually expressed in a non-redundant way.

A 3D rigid body transformation $T \in SE(3)$ denotes rotation and translation in 3D:

$$T = \begin{pmatrix} R & t \\ 0 & 1 \end{pmatrix} \quad \text{with} \quad R \in SO(3) \quad \text{and} \quad t \in \mathbb{R}^3$$

The optimisation purpose in image alignment is to find the transformation $T$ in each time step, i.e. $T$ is regarded as the camera pose. A minimal representation for camera pose is better for optimisation purpose. The Lie algebra $se(3)$ corresponding to the tangent space of $SE(3)$ at the identity is used as the minimal representation. The algebra element is called twisted coordinates $\xi = [\omega^T \nu^T] \in \mathbb{R}^6$. The map from Lie algebra $se(3)$ to Lie group $SE(3)$ is the exponential map $T(\xi) = \exp(\psi(\xi))$ and its inverse map is the logarithm map $\psi(\xi) = \log T(\xi)$, where $\psi(\xi)$ is the wedge operator,

$$\psi(\xi) = \begin{pmatrix} \omega \times & \nu \\ 0 & 1 \end{pmatrix}$$

A 3D point with homogeneous vector $c_p_u$ in the camera frame maps to the image coordinate $u$ via the pinhole camera projection model:

$$u = \pi(c_p_u) = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} + \frac{f_u}{f_v} \begin{pmatrix} f_u & 0 \\ 0 & f_v \end{pmatrix} \begin{pmatrix} \frac{z}{s} \\ 1 \end{pmatrix}$$

where $u_0, v_0$ and $f_u, f_v$ are the principal point and focal length, respectively, representing camera intrinsic parameters which can be calibrated beforehand. Given a depth information $d_u$ for a point $u$, the 3D point in the camera frame can be recovered from an image coordinate:

$$c_p_u = \pi^{-1}(u, d_u)$$

III. VISUAL ENTROPY

A. Information entropy

Generally, the entropy is defined for two different levels of description of a given system: the macroscopic state of the system is defined by a distribution on the microstates that are accessible to a system in the course of its thermal fluctuations. At one of these levels, the entropy $S$ is given by the Gibbs' entropy formula, $S = -k_B \sum p_i \ln p_i$, where $p_i$ is the probability that it occurs during the system’s fluctuations and the quantity $k_B$ is a physical constant known as Boltzmann’s constant.

In information theory, entropy $H(X)$ defines the theoretical number of bits needed to encode a random variable $X$. This random variable could stand for an event, sample or character drawn from a distribution or data stream. Here, for our camera pose estimation, we consider the images or regions of interest as the random variables with possible valuates as the intensity of pixels whose possibility can be calculated from their histogram. Thus, if a discrete random variable $X$ with possible values $\{x_1, \cdots, x_n\},$ then the Shannon entropy $H(X)$ is given by:

$$H(X) = \sum_{i=1}^{n} P(x_i) \log_2 (P(x_i))$$

where $H(X) = - \sum_{i=1}^{n} P(x_i) \log_2 (P(x_i))$ represents the self-information. For mathematical completeness, $0 \log_2(0) = 0$ should be introduced. Intuitively, the more values $x$ are equally probable the bigger entropy $H(X)$ is, and the entropy reaches maximum: $H(p_1, \cdots, p_n) = H(1/n, \cdots, 1/n) = \log_2(n)$. While pose estimation matters sequential images rather than only one image itself, building up the relation between two random variables becomes quite important.

B. Joint entropy

Joint entropy $H(X, Y)$ describes the uncertainty associated with a set of variables, defining the theoretical number of bits needed to encode a joint system of two random variables $X, Y$ with possible values $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$ respectively.

$$H(X, Y) = - \sum_{i=1}^{n} \sum_{j=1}^{m} P(x_i, y_j) \log_2 (P(x_i, y_j))$$

In the above expression, $P(x, y)$ is the joint probability of these values occurring together.

The joint entropy is bounded by $\max (H(X), H(Y)) \leq H(X, Y) \leq H(X) + H(Y)$, the second equality happens if and only if $X$ and $Y$ are statically independent, while the first inequality becomes equal when $Y \subseteq X$ or $X \subseteq Y$, which means $X$ or $Y$ can fully represent the other. In this case, a alignment problem can be viewed as finding the minimum of the joint entropy, thus making the two set of variables included by one another in a large extent.
However, it is not such simple for us just using the joint entropy to construct a cost function for pose estimation. In practical, adding variable \( Y \) into the system only increases the joint entropy \( H(X, Y) \) but adding a variable set, which is exactly equal to original set \( X \), i.e. \( X = Y \), or only a constant set (probability is zero) would not add variability to the system. In the constant case, the set \( Y \) obviously can not express \( X \), thus the alignment fails. However, the conception of mutual information has been proposed to solve this issue [11].

### C. Mutual information

The mutual information, also formally called transinformation of two random variables, is a measure of the variables’ mutual dependence. For two random variables \( X \) and \( Y \), their mutual information is given by the following equation:

\[
I(X, Y) = H(X) + H(Y) - H(X, Y).
\]

Substituting equation (3) and (4), it yields to:

\[
I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log \left( \frac{P(x, y)}{P(x)P(y)} \right)
\]

where \( P(x, y) \) is the joint probability distribution function of \( X \) and \( Y \), and \( P(x) \) and \( P(y) \) are the marginal probability distribution functions of \( X \) and \( Y \) respectively.

As we can see in equation (5), mutual information integrates the individual entropy of each variable and their joint entropy. If the random variables \( X \) and \( Y \) are independent, then \( I(X, Y) = 0 \); if they totally depends on each other, say \( X = Y \), then \( I(X, Y) = H(X) = H(Y) \). It can be easily checked that even for the set \( Y \) including most constant outliers, \( I(X, Y) = 0 \) due to \( H(Y) = 0 \), then the alignment will fail. Thus, the bigger the mutual information is, the better the aligned results would be. Compare to the sum of squared differences (SSD), this alignment method does not need to find the linear relation between two signals [12].

### IV. POSE ESTIMATION METHODS

#### A. Pose estimation based on mutual information

The goal of visual-based estimation is to find the current pose through image alignment. Here the general random variables \( X \) and \( Y \) become the images, and the pixel intensity becomes possible value within image domain.

As stated in section II, a rigid body transformation \( \xi \) denotes rotation and translation and a point in 3D space with homogeneous vector \( \mathbf{p}_u \) in the camera frame maps to the image coordinate \( u \) via the camera projection. With this transformation \( \xi \), the mutual information of two images is given by:

\[
I(X, Y) = \sum_i \sum_j P(i, j, \xi) \log \left( \frac{P(i, j, \xi)}{P(i, \xi)P(j)} \right)
\]

where \( i \) and \( j \) are pixel intensities in images \( I_c \) and \( I_c' \), respectively. \( P(i, \xi) \) and \( P(j) \) are respectively the probability of the intensity \( i \) and \( j \) in the images, and \( P(i, j, \xi) \) is the joint probability of two intensities. They can be computed as a normalized histogram:

\[
P(i, \xi) = \frac{1}{N} \sum_u \Phi \left( i - \frac{j}{\lambda} I_c(u, \xi) \right)
\]

\[
P(j) = \frac{1}{N} \sum_u \Phi \left( j - \frac{i}{\lambda} I_c'(u) \right)
\]

\[
P(i, j, \xi) = \frac{1}{N} \sum_u \Phi \left( i - \frac{j}{\lambda} I_c(u, \xi) \right) \Phi \left( j - \frac{i}{\lambda} I_c'(u) \right)
\]

where \( N \) is the number of pixels in selected subset, \( \lambda \) is the gray levels of the original image and \( \lambda' \) is a desired value space with smaller scale to increase the computational efficiency and robustness [10]. \( \Phi(\cdot) \) is B-spline function [13] with the character of unit result for integral operation, thus getting rid of renormalising.

Having the mutual information between images, we can acquire the best pose estimation through an optimization problem as

\[
\xi^* = \arg \min_{\xi} (-I(I_c(\xi), I_c')).
\]

This can be solved by Levenberg-Marquardt [14],

\[
\xi = -\alpha (H_\xi + \beta |H_\xi|d)^{-1} G_\xi
\]

where \( H_\xi \) and \( G_\xi \) are the Gradient and Hessian matrix of equation (9) with the parameters \( \alpha \) and \( \beta \). The Gradient can be computed as

\[
G_\xi = -\frac{\partial I(I_c(\xi), I_c')}{\partial \xi} = -\sum_{i,j} \frac{\partial P(i, j)}{\partial \xi} \left( 1 + \log \left( \frac{P(i, j)}{P(i)} \right) \right)
\]

with the joint probability that can be calculated further by chain rule in equation (8),

\[
\frac{\partial P(i, j)}{\partial \xi} = -\frac{1}{N} \sum_u \frac{\lambda' \Phi \partial I_c \partial u \Phi \left( j - \frac{i}{\lambda} I_c' \right)}{\partial \xi}
\]

where \( \alpha \) is the image gradient and \( \frac{\partial u}{\partial \xi} \) is the interaction matrix at point \( u \), given by a the projection in section II:

\[
\frac{\partial u}{\partial \xi} = \begin{pmatrix} -1/d_u & 0 & u/d_u & u v & -(1 + u^2) v \\ 0 & -1/d_u & v/d_u & 1 + v^2 & -w u & -u \end{pmatrix}
\]

The depth value \( d_u \) above is typically estimated through a mapping process in VO or SLAM. Having the Gradient matrix the Hessian is approximately given by omitting a higher order term.
B. Pose estimation based on entropy-like cost function

For a general non-linear system with parameter state vector \( x(t) \) and other known inputs \( u \), an observation \( z_i \), \( i \in 1,2,...,N \) can be written as

\[
z_i = h(u, x(t)) + v_i
\]

where \( v_i \) refers to the process noise in each observation. A relative squared residuals \( e_i \) is defined as

\[
e_i = \frac{\tilde{z}_i^2}{\sum_{j=1}^{N} \tilde{z}_j^2}
\]

where \( \tilde{z} \) is the estimation of the observation containing the unknown estimation \( \hat{x} \) and the relative squared residuals as the normalised factor can be regarded as some sort of the probability. Thus, we can write a normalised entropy-like cost function as follows:

\[
H = -\frac{1}{\log N} \sum_{i=1}^{N} e_i \log e_i
\]

with \( H = 0 \) if \( \sum_{i=1}^{N} e_i^2 = 0 \). The zero condition represents an exact matching and also completes the definition.

An intuitive understanding of this cost function is that the best matching leading to less residuals and more outliers would make large residuals. In order to increase the robustness, the relative squared residuals are often wrapped by Huber-like function \[15\] to limit the upper bound of the denominator.

C. Feature based pose estimation

Here for comparison, we briefly list a weighted cost function to find the camera pose and adjust the features' position at the same time. Normally, the re-projection error is defined as the difference between a measurement \( z \) and its estimate \( \hat{z}(\xi, \hat{p}_u) \):

\[
\tilde{z} = z - \hat{z}(\xi, \hat{p}_u)
\]

The cost function \( \eta(\xi, \hat{p}_u) \) is the sum of all squared errors \( \tilde{z} \) with a weighting matrix \( W \):

\[
\eta(\xi, \hat{p}_u) = \sum_{i=1}^{n} \sum_{j=1}^{m} \tilde{z}_{i,j}^Tw_{i,j}\tilde{z}_{i,j}
\]

where \( j \) from 1 to \( m \) is the index of points within a frame, and \( i \) is the number of frames indexing a set with size \( n \).

Then, the optimisation problem is defined as

\[
\xi, \hat{p}_u = \arg \min_{\xi, \hat{p}_u} \eta(\xi, \hat{p}_u)
\]

V. RESULTS

This section presents some results of the process in pose estimation using different expressions mentioned above. We focus on the aspects of environmental condition change, average computational time before optimisation calculation, and the threshold of mutual information. The results indicate that with an equivalent accuracy, the entropy based methods enjoy the merits of robustness and computational efficiency and it is worth to further develop for applications in high flexible platforms like micro aerial vehicles (MAVs). The results were performed by applying consecutive images taken from our office environment. The frame rate is 30 Hz with the resolution of 640x480 and the lens is of 120° view angle with fish-eye distortion. All computational work was done in a laptop with Intel i7-5600U CPU and 8G RAM. Some sample sequential images in our trail are shown in Figure 2.

A. Robust to illumination

Some experiments took place under illumination change environment, for example turning off the lights during one trial. This simulates the situation which happens usually in real-time application: the camera moving from indoor to outdoor or merely used in dark environment. At our presented scene in Figure 3, there are four consecutive images including two under less illumination conditions. The matching lines between features show that after changing the illumination, matching pairs become less in number (decreasing 20 % as previous), limited in area (focusing on right area other than full image range) and more in mistake (outliers). The results indicate that in the direct-based expression, the pixel intensity are expressed in probability according to the histogram of whole image, which means except for the environment of totally dark, changing the illumination of the whole image would make less impact on the matching results compared to feature-based methods.
**B. Stable in mutual information**

In order to show the availability and trend of mutual information, we formed two image sequences including 10 frames, which down-sampled from original sequence at 1/2 and 1/10 of original rate. In Figure 4, we can obviously view that values of mutual information wave around 2 between every other two frames (the diagonal line starting from B in Figure 4(a)), and the values jolt around 1 between every ten frames (the diagonal line starting from F in Figure 4(a) and B in Figure 4(b)). The values drop from average 5.8 to 2 dramatically between close frames but slightly between distant frames (four of five intervals). Thus, based on this value, a distance threshold as a constraint for motion optimisation can be set. The stable variance of mutual information also reveals the robustness character of entropy-based evaluation although the environment changes significantly.

**C. Efficient in computation**

Figure 5 shows the computational times of mutual information based on direct image information and standard SIFT process. The average computational time for mutual information for two trials are 0.0594s and 0.0440s, but for feature process they are 20.8772s and 18.5485s. The processes were...
This paper introduced the entropy-based pose estimation methods using visual information from a monocular camera. Stemming from information theory, entropy-like methods by natural have the robustness due to the use of probability, thus widely used to construct the cost function for optimisation. In another aspect, different visual observations and expressions under the same conception of entropy will trigger dramatically different character in performance. Specifically, entropy-based optimisation depending on direct data from images or features extracted through a process will cost dramatically different computational consumption in whole algorithm. The entropy-based method using mutual information also shows the advantage in illumination variable environment. Through the experiments, we can see the huge potential in applying the entropy-based pose estimation methods in highly agile platforms like MAVs. In the future, we will further improve and analyse these methods and target at more real-time applications.

ACKNOWLEDGMENT

The first author has been financially supported by scholarship from China Scholarship Council.

REFERENCES


