Scene Matching by Spatial Relationships

Ozy Sjahputera, James M. Keller, Fellow, IEEE, and Pascal Matsakis

Abstract—Scene matching is the process of recognizing two images as different views of the same scene captured using different sensor poses, and/or different types of sensors. In this work, each image contains the same objects and sensor pose parameters are not known. The spatial relationships among objects in the image, calculated using the histogram of forces (F-histogram) method, are used as matching elements. The degree of matching between two matching elements is calculated by comparing their F-histogram representations. Various geometric transformations are applied to the F-histograms during the comparison process to maximize the histogram similarity measure and to estimate the sensor pose parameters. The histogram similarity measure and the estimated sensor pose parameters are used as features in finding the best histogram correspondence map that matches the two images.

Index Terms—Histogram of Forces, Histogram Comparison, Histogram Transformation, Scene Matching, Spatial relations

I. INTRODUCTION

Scene matching refers to the ability to locate or match a region of an image representing a view of a scene with a corresponding region of another view of the same scene often taken under different sensor pose geometry, or different type of sensor. Points, edges, end points of edges, lines, line intersections, and color are common matching elements used in various techniques for scene matching. Variations in sensor pose used to capture the two images can cause corresponding matching elements in both images to have different position and orientation. Geometric transformations applied to the images may help in minimizing these differences, allowing a higher degree of matching when the two images, or regions of the images, are correlated. In one of the early work, Wong [1] proposed two geometric transformations: perspective-based and polynomial estimation approach. The former approach required the sensor pose parameters to be known. The latter approach estimated the transformation using pairs or corresponding points (matching elements) available in both images. Wong also used edge information as matching elements in [2]. Instead of applying the geometric transformation, he proposed a hierarchical sequential search to find the image region with maximum cross correlation [3]. The VLSI design for this algorithm was proposed in [4] and implemented in [5]. The method is flexible enough to identify a scene viewed at different times, or under different seasonal conditions, but is designed more for scenes taken from the same angle. Boland et al. in [6] used the gradient quantization to produce the edge map for scene matching. The matching performance is improved by optimizing the threshold for gradient quantization. Shi et al. [7] proposed an algorithm based on edge enhancement and Fourier phase correlation under the assumption that the considered images have very similar edge orientations.

Han and Park [8] used epipolar geometry for contour matching across views. The degree of matching was assessed using a correlation-based technique by taking into account the epipolar constraints. In [9], Terefa and Harada used corner points as matching elements to recover the epipolar geometry of two uncalibrated images of the same scene. Schmid and Zisserman [10] used both the line information and epipolar geometry to automatically match lines across views. The epipolar geometry reduced the search complexity by restricting line segments need to be considered for matching.

Spatial relations refer to the study of relative position between objects in an image. The use of angles to define directional relative position of an object was proposed by Krishnapuram et al. [11]. This concept was used in the angle histogram method introduced by Miyajima and Ralescu [12]. Most recently, Matsakis and Wendling proposed the force histogram method [13] that generalized the angle histogram. Mathematical morphology had also been used to define spatial relationships. Gader [14] fuzzified the standard method for computing binary morphology using the extension principle. He then used the fuzzified method to compute fuzzy spatial relations between objects in digital images. More recent use of mathematical morphology in spatial relations study can be found in [15].

In [16], Keller et al. used linguistic descriptions of spatial relations as matching elements in scene matching problems. They used the method proposed in [14] to calculate the spatial relationships. Matsakis and Wendling [17] used the force histograms to match two object pairs based on their spatial relations similarity. This approach was extended to scene matching by Sjahputera et al. in [18]. This work was further improved in [19] by introducing a better histogram matching algorithm that used various force histogram transformations to optimize the histogram similarity and to estimate the sensor pose parameters (tilt, azimuth, scaling factor).
II. SCENE REPRESENTATION USING SPATIAL RELATIONS

A. Histogram of Forces (F-Histograms)

In the histogram of angles $A_0$ method [12], image objects are treated as collections of pixels, hence $A_0$ can be assessed on raster data only. $A_0$ associated with any pair $(A,B)$ of crisp and digitized objects is a function $\text{Ang}^{AB}$ from $R$ into $N$. For any direction $\theta$, the value $\text{Ang}^{AB}(\theta)$ is the number of pixel pairs $(p,q)$ belonging to $A \times B$ such that $p$ is in direction $\theta$ of $q$. In [17], Matsakis and Wendling introduced the histogram of forces (F) that generalizes and supersedes $A_0$. $F$-histograms can be calculated from both crisp and fuzzy objects represented as either raster or vector data. The $F$-histogram associated with $(A,B)$ is a function $F^{AB}$ from $R$ into $R^+$. Like $\text{Ang}^{AB}$, this function represents the relative position of $A$ with regard to $B$. For any direction $\theta$, the value $F^{AB}(\theta)$ is the total weight of the arguments that support the proposition “$A$ is in direction $\theta$ of $B$”. More precisely, it is the scalar resultant of elementary forces exerted by the points of $A$ on those of $B$, and each tends to move $B$ in direction $\theta$. Different types of $F$–histogram can be obtained by using different definitions of the elementary forces exerted by points in $A$ to those in $B$. For this work, we define the elementary forces as $d^{-\theta}$, where $d$ represents the distance between the points considered and $r$ is a real, then $F$ is denoted $F_r$. The $F_2$–histogram (gravitational forces) and $F_0$–histogram (constant forces) have very different and very interesting characteristics. The latter, very similar to $A_0$, gives a global view of the situation. It considers the closest parts and the farthest parts of the objects equally, whereas $F_2$–histogram focuses on the closest parts.

![Fig. 1. The notion of histogram of forces: (a) elementary forces in (A,B) tend to move B in direction $\theta$, (b) $F_r^{AB}$ (constant forces), (c) $F_2^{AB}$ (gravitational).](image)

B. $F$-histograms as Object Pairs Representations

Let $S$ and $S'$ be different views of the same scene. For this work, we assume that $S$ and $S'$ contain the same number of objects ($N$) and for each object in $S$ there exists a corresponding object in $S'$, i.e. there exists a bijective mapping between the objects in $S$ and those in $S'$. We consider all possible unique object pairs $(A,B)$ in each view such that 1) $A \neq B$, 2) the opposite pair is excluded (i.e., if $(A,B)$ is considered, then $(B,A)$ is excluded since it is a $\pi$-rotation of $(A,B)$). Thus, the number of unique object pairs generated from $N$ objects is given by $(N^2-N)/2$. An $F_r$–histogram is calculated for each object pair. $F_0$ is used based on the analysis given in [19]. The bijective mapping exists between the $F_r$–histograms of $S$ and those of $S'$, i.e. for each $F_r^{AB}$ in $S$ there exists $F_0^{AB}$ in $S'$ such that $A=A'$ and $B=B'$, or $A=B'$ and $B=A'$. An example for $N=3$ is given in Fig. 2.

![Fig. 2. Object pairs are represented using $F$-histograms. The correct bijective $F$-histogram mapping is shown by the lines connecting every $F$-histogram in (a) to its correspondence in (b).](image)

III. MATCHING METHODS

A. F-histogram Matching

Given $F_r^{XY}$ in $S$ we need to find $F_r^{XY'}$ in $S'$ such that the histogram similarity measure $\mu(F_r^{XY}, F_r^{XY'})$ is maximum for all object pairs $(X',Y')$ in $S'$. Three sensor pose parameters are considered: tilt angle, azimuth orientation, and scaling factor. The tilt and azimuth are illustrated in Fig. 3(a). The swing is assumed to be always $+180^\circ$, i.e., a plumb line that intersects the principal ray is mapped to the Y-axis in the image plane. Differences in sensor pose parameters can cause the same object to appear differently in $S$ and $S'$, i.e. $\mu(F_r^{AB}, F_r^{A'B'})$ may not yield the highest similarity even though $A=A'$ and $B=B'$. Traditionally, this condition is handled by applying geometric transformations on the images. Consider the object pair $(A_0,B_0)$ in $S_0$ and $(A'_0,B'_0)$ in $S'_0$ as shown in Fig. 4. Note how the two object pairs look very different initially. A series of transformations is applied to minimize the differences in sensor parameters. The final outputs, $(A_4,B_4)$ in $S_4$ and $(A'_4,B'_4)$ in $S'_4$, are identical in every respect, i.e. correlating $S_4$ and $S'_4$ will yield a perfect match (assuming vector data are used). In this work, the geometric transformations are adapted to operate on the $F_r$-histogram representation of the objects rather than the objects themselves. The transformations are derived from $F_r$-histogram properties detailed in [17] and [19]. The $F_r$-histogram matching algorithm used here was first proposed and detailed in [19].
Given the image \( S_0 \) and \( S'_0 \), we calculate \( F_{r,A_1B_1}^{S_0} \) and \( F_{r,A_1'B_1'}^{S'_0} \) respectively. Let \( \cos \theta > 0 \) and \( \cos \theta > 0 \) be the tilt angle on the captured image. The projection of a segment as seen from some tilt angle \( \cos \theta > 0 \) is shorter by the factor of \( \cos(t) \) compared to the real size.

Let \( n_0 \) and \( n_1 \) be the number of directions considered for \( F_{r,A_1B_1}^{S_0} \) and \( F_{r,A_1'B_1'}^{S'_0} \), respectively. For \( t = [0^\circ, 60^\circ] \), we need to have \( n_0=2n_1 \) to avoid sub-sampling effect from occurring in Equation (1) [19]. The same transformation with \( k'=1/\cos(t') \) is used to obtain \( F_{r,A_1'B_1'}^{S'_0} \) from \( F_{r,A_1B_1}^{S_0} \). See [19] for details.

2) Scaling Factor: Equalizing the scaling factor of \( S_0 \) to that of \( S'_1 \) results in \( S_2 \), i.e. \( (A_2,B_2) \) is equal in size as \( (A'_1,B'_1) \). \( F_{r,A_1B_1}^{S_0} \) can be derived directly from \( F_{r,A_1'B_1'}^{S'_0} \) as follows:

\[
F_{r,A_1B_1}^{S_0}(\theta) = \ell^3 r F_{r,A_1'B_1'}^{S'_0}(\theta)
\]

\( \ell \) is the dilation ratio given by \( \ell = (m'/m)^{3-r} \) where \( m' \) and \( m \) are the means of \( F_{r,A_1'B_1'}^{S'_0} \) and \( F_{r,A_1B_1}^{S_0} \), respectively.

3) Azimuth: Except for the orientation, the pair \( (A_3,B_3) \) is now identical in shape and size to \( (A'_1,B'_1) \). Let \( \rho \) be the difference in azimuth orientation between \( S_2 \) and \( S'_1 \):

\[
\rho = \begin{cases} 
  c' - c & \text{if } c' \geq c \\
  (c' - c) + 360^\circ & \text{otherwise}
\end{cases}
\]

where \( c' \) and \( c \) represents the X-axis coordinate of the centroids of \( F_{r,A_1B_1}^{S_0} \) and \( F_{r,A_1'B_1'}^{S'_0} \) respectively, where \( c'=c+\rho \). Rotating \( S_2 \) by \( \rho \) results in \( S_3 \) and \( S_3 \) matches \( S'_1 \). Thus, \( F_{r,A_1B_1}^{S_0} \) can be derived from \( F_{r,A_1'B_1'}^{S'_0} \) as follows:

\[
F_{r,A_1B_1}^{S_0}(\theta) = F_{r,A_1'B_1'}^{S'_0}(\theta - \rho)
\]

4) Translation: \( F_{r}-histogram \) is independent of translation. Thus, \( F_{r,A_1B_1}^{S_0} = F_{r,A_1'B_1'}^{S'_0} = F_{r,A_2B_2}^{S_3} \).

5) Histogram Similarity Measure: Given two histograms, \( h_1 \) and \( h_2 \), their degree of similarity (in \([0,1]\) range) is defined as:

\[
\mu(h_1,h_2) = \frac{\sum \min(h_1(\theta),h_2(\theta))}{\sum \max(h_1(\theta),h_2(\theta))}
\]

6) F-Histogram Matching Algorithm: We need to find the 4-tuple \( (\rho, \ell, t, t') \) that maximizes \( \mu \). These values represent the sensor pose parameters and in many cases they are not known. \( \ell \) and \( \rho \) can be estimated as shown in step 2) and 3). Marjamaa et al. [20] estimated the tilt angle from the LADAR range information of the image. However, range information is not available for images captured using optical sensor. Therefore, to maintain generality, we simply search for the tilt angles, \( t \) and \( t' \) since they cannot be estimated directly from the
histograms. The algorithm is given as follows:

0. $\alpha \leftarrow 0$.
1. Compute $F^{r_0}_{r_0}$ (directly from $(A_0, B_0)$) 
2. Compute $F^{r_0}_{r_0}$ (directly from $(A_0, B_0)$) 
3. For each $t'$
   3.1. Compute $F^{r, A_0}_{r, A_0}$ from $F^{r, A_0}_{r, A_0}$ (Equation (1))
   3.2. For each $t$
       3.2.1. Compute $F^{r, A_0}_{r, A_0}$ from $F^{r, A_0}_{r, A_0}$ (Equation (1))
       3.2.2. Compute $\ell$ from $F^{r, A_0}_{r, A_0}$ and $F^{r, A_0}_{r, A_0}$
       3.2.3. Using $\ell$ compute $F^{r, A_0}_{r, A_0}$ from $F^{r, A_0}_{r, A_0}$
             (Equation (3))
       3.2.4. Compute $\rho$ from $F^{r, A_0}_{r, A_0}$ and $F^{r, A_0}_{r, A_0}$
             (Equation (4))
       3.2.5. Using $\rho$ compute $F^{r, A_0}_{r, A_0}$ from $F^{r, A_0}_{r, A_0}$
             (Equation (5))
       3.2.6. If $\sigma < \mu(F^{r, A_0}_{r, A_0}, F^{r, A_0}_{r, A_0})$
            then $\alpha \leftarrow \mu(F^{r, A_0}_{r, A_0}, F^{r, A_0}_{r, A_0})$

   $R \leftarrow \rho$, $L \leftarrow \ell$, $T \leftarrow t$ and $T' \leftarrow t'$.

B. Scene Matching

Let $H = \{h_0, \ldots, h_i, \ldots, h_{i,\text{max}}\}$ and $H' = \{h'_0, \ldots, h'_i, \ldots, h'_{i,\text{max}}\}$ be the sets of all F-histograms in $S$ and $S'$. Since $S$ and $S'$ contains the same N objects, $|H| = |H'| = (N^2 - N)/2$, and $i_{\text{max}} = j_{\text{max}} = ((N^2-2-N)/2)-1$. The number of possible bijective mappings between $H$ and $H'$ is given by $|H|!$. However, only one of these mappings correctly match each histogram in $H$ to its correspondence in $H'$. Let $M = \{m^0, \ldots, m^i, \ldots, m^{r, \text{max}}\}$ be a set of all possible bijective mappings from $H$ to $H'$, $r_{\text{max}} = |H|!$. $m^i$ is a bijective mapping defined as $m^i = \{m^0_i, \ldots, m^j_i, \ldots, m^{r, \text{max}}_i\}$ where $m^i$ contains the histogram number in $H'$ being matched to the $i$-th histogram $h_i$. An example of a bijective mapping $m^i$ is shown in Fig. 5. Is it a correct mapping?

![Diagram of bijective mapping](image)

Fig. 5. An example of a bijective mapping from (a) to (b). Each histogram matching produces 5-tuple output $(\sigma_t, R_t, L_t, T_t)$, $T_{r'}$. All matching outputs are aggregated to produce the matching degree between (a) and (b).

Each histogram matching specified in $m^i$ is processed using the matching algorithm given in Section III.A.6, producing the outputs $(\sigma^r_t, R^r_t, L^r_t, T^r_t, T_{r'}.r)$. Thus, for each mapping $m^i$ we produce $i_{\text{max}}+1$ 5-tuple outputs. In theory, if the mapping $m^i$ is correct, the estimated sensor pose parameters obtained during histogram matching $(R^r_t, L^r_t, T^r_t, T_{r'}^r)$ should be similar in value across all $m^i \in M$. Hence, the outputs $R^r_t, L^r_t, T^r_t, T_{r'}^r$ are aggregated by measuring the similarity of their values rather than the magnitudes. The aggregations are performed as follows (similarity measure):

$$\sigma^r = \frac{1}{i_{\text{max}}+1} \sum_{t=0}^{i_{\text{max}}+1} \sigma^r_t$$

$$R^r = 1 - \frac{\sum_{i=0}^{i_{\text{max+1}}} \sum_{j=0}^{i_{\text{max}}} \text{rotDist}(R^r_i - R^r_j)}{i_{\text{max+1}}(i_{\text{max+1}}-1)/2}$$

where $\text{rotDist}(R^r_i - R^r_j) = \begin{cases} |R^r_i - R^r_j| & \text{if } |R^r_i - R^r_j| < 180^\circ \\ \text{otherwise} & \end{cases}$ is the rotational distance between two angles.

$$\ell^r = \frac{1}{i_{\text{max+1}}} \sum_{i=0}^{i_{\text{max+1}}} \|\log_{10}(L^r_i) - \mu^r_i\|$$

where $\mu^r_i = \frac{1}{i_{\text{max+1}}} \sum_{i=0}^{i_{\text{max+1}}} \log_{10}(L^r_i)$. $\ell^r$ is the variance of $L^r_i$. The similarity measure $L^r$ is scaled to [0,1] with respect to all $\ell^r$ belonging to all mappings $m^i$ in $M$, i.e. the mapping $m^i$ that produces the smallest $\ell^r$ will have $L^r = 1$.

$$T_{r'}^r = 1 - \frac{1}{360^\circ} \sum_{i=0}^{i_{\text{max+1}}} (\sum_{j=0}^{i_{\text{max}}} (|T^r_i - \mu^r_i| + |T_{r'}^r - \mu_{r'}^r|))$$

where $\mu^r_i$ and $\mu_{r'}^r$ are the averages of $T^r_i$ and $T_{r'}^r$. Note that $\sigma^r$, $R^r$, $L^r$, and $T^r$ are all in [0,1].

Let $\Sigma$ be the degree of matching between $S$ and $S'$ using the mapping $m^i$. We compute $\Sigma$ using two methods: simple averaging (ave$\Sigma$) and choquet integral (cho$\Sigma$). Using exhaustive search, we choose the histogram mapping $m^i$ with maximum $\Sigma$ as our best mapping to match $S$ and $S'$.

IV. EXPERIMENTAL RESULTS

A. Experimental Data

Our data is divided into 3 groups, all of which contain images captured from real world scenes. Group1 contains 7 views of a scene with 4 objects. Group1 contains 3 views of a scene with 5 objects; 2 of the objects are identical. Group3 contains 6 views of 5 identical objects. Group1 and Group2 are extracted from a set of LADAR images supplied by NAWC (Naval Air Warfare Center), China Lake, Ca. The objects are manually segmented. The objects in Group3 are segmented automatically by a simple thresholding method. In all three groups, the objects are labeled consistently across all views. We generate all possible unique view pairs within each group;
each pair contains different views. This way, we have 21 view pairs for Group1, 3 view pairs for Group2, and 15 view pairs for Group3, giving the total of 39 view pairs. Our experimental objective is to find the correct histogram mapping for each view pair using the methods described in Section III. Currently, the best mapping is found by exhaustive search. The histogram matching algorithm given in Section III.A.6. is executed by varying the values of $t$ and $t'$ from 0° to 60° with a 5° increment.

![Fig. 6. Examples of real world scenes used in the experiments.](image)

**B. Matching Results**

We first match the 39 view pairs using $\sigma^r$, $R^r$, $L^r$, and $T^r$ alone. The individual performances are used as densities for calculating $\text{cho}^r\Sigma'$ with $\lambda$ fuzzy measure. We found the individual performance of $\sigma^r$, $R^r$, $L^r$, and $T^r$ at 0.1, 0.41, 0.03, and 0.03, respectively. The number of view pairs that are matched successfully using all four features $\sigma^r$, $R^r$, $L^r$, and $T^r$ with $\text{ave}^r\Sigma'$ and $\text{cho}^r\Sigma'$ are presented in Table 1.

![Fig. 7 Error case 1 (cho$\Sigma'$): except for the $F^{12}_r$ and $F^{04}_r$ in S, all other histograms are matched correctly. The dark arrows indicate the incorrect matching. The grey arrows indicate the actual matching.](image)

**C. Error Analysis**

For the error case shown in Fig. 8, the histograms $F^{01}_r$ and $F^{02}_r$ are matched incorrectly. All other histograms are matched correctly. Objects 1 and 2 (1' and 2') represent two identical buildings. The object pair (0,2) in S is supposed to match with (0',2') in S'. However, in sensor orientation, (0,2) appears to be more similar (in terms of spatial relationship) to (0',1') than it is to (0',2'). This explains the error in histogram mapping. As in the previous error case, the correct histogram mapping for this view pair experiment is also found with the 2nd highest value of $\text{cho}^r\Sigma'$.

The two error cases above show that the “error” produced by our system to be non-catastrophic, i.e. the correct answer is very close to the top, hence the chance of “recovering” the correct mapping using some post processing (such as using object information) is very good.

![Fig. 8. Error case 1 (ave$\Sigma'$): except for the $F^{01}_r$ and $F^{02}_r$, all other histograms are matched correctly. The dark arrows indicate the incorrect matching. The grey arrows indicate the actual matching.](image)
V. CONCLUSIONS AND FUTURE WORK

In this paper, we proposed a new approach to scene matching using spatial relationship information among objects in the scene as the matching elements. The histogram of forces method is selected to represent the spatial relationship between objects in the scene. This method offers solid mathematical foundations and allow the processing of both raster and vector data. Finding the best match for an object pair can be implemented as the matching operation of two F-histograms. For this, we propose an F-histogram matching algorithm that maximizes the histogram similarity measure. The algorithm follows the conventional concept of applying geometric transformations to minimize the effect caused by differences in sensor pose parameters. In our approach, the geometric transformations are not applied to the images. Instead, we use the properties of F-histogram method to derive equivalent geometric transformations to operate directly on the F-histograms. The matching algorithm also provides the estimates of sensor pose parameters for the two views.

In this work, each image is represented by a set of F-histograms generated from all possible (unique) object pairs contained in the image. The process of matching two scenes that are known to contain the same objects (but viewed from different sensor orientations) is reduced to the process of finding a histogram mapping between the two images with the maximum scene degree of matching. The estimated sensor pose parameters produced by the F-histogram matching algorithm are consistent in values when the correct histogram mapping is used. Hence, the variance of the pose parameters estimates are used to calculate the degree of matching associated with a specific histogram mapping. We showed that our approach is able to produce good matching performance (90%) on 39 view pairs taken from various real world scenes. We also showed how the use of spatial relationship information in scene matching can be successful even on cases where object recognition alone will not be able to succeed (see Group3 where all objects in the scenes are identical).

Exhaustive search to find the best histogram mapping is too costly for practical application. The number of possible solution increases factorially as the number of objects in the image increases. Hence, we plan to develop a better method to search for the best histogram mapping. We also plan to look at cases of partial matching where only several of the objects contained in the images actually match.

REFERENCES