MULTIMODAL TEMPLATE MATCHING BASED ON GRADIENT AND MUTUAL INFORMATION USING SCALE-SPACE

Fernando Barrera[⋆]*, Felipe Lumbreras*[⋆‡]*, and Angel D. Sappa*[⋆]

Computer Vision Center[⋆] and Computer Science Department[‡]
Universitat Autònoma de Barcelona
Edifici O, Campus UAB, 08193 Bellaterra, Spain

ABSTRACT

This paper presents the combined use of gradient and mutual information for infrared and intensity templates matching. We propose to joint: (i) feature matching in a multiresolution context and (ii) information propagation through scale-space representations. Our method consists in combining mutual information with a shape descriptor based on gradient, and propagate them following a coarse-to-fine strategy. The main contributions of this work are: to offer a theoretical formulation towards a multimodal stereo matching; to show that gradient and mutual information can be reinforced while they are propagated between consecutive levels; and to show that they are valid cost functions in multimodal template matchings. Comparisons are presented showing the improvements and viability of the proposed approach.

Index Terms— Mutual information, Pattern matching, and Infrared imaging.

1. INTRODUCTION

The coexistence of infrared cameras with other sensors has opened new perspectives for the development of multimodal systems. One of the challenges is to find the best way to fuse all this information in a useful representation. In the current work the matching of images from different spectral bands is considered. These images are provided by a calibrated stereo rig, which consists of two cameras capable of measuring emissions in visible/infrared spectrum.

The literature on multimodal template matching can be broadly divided into entropy-based methods and feature-based methods. In the current work a hybrid approach is proposed. It exploits mutual information and gradient information, in scale-space representations. Mutual information is a concept derived from information theory. It measures the amount of information that one random variable contains about another. It is a powerful concept in situations where no prior relationships between the data are known. The previous property makes mutual information the ideal tool to address problems involving signals without an apparent relationship.

Viola intuitively introduces mutual information as a measure of alignment between images and 3D models [1]; next, it was formalized in [2], only for images. The importance of these early contributions was to exploit properties of mutual information in the field of multimodal image processing, showing its usefulness. Just few years later, a cost function (or dis/similarity function) for stereo vision [3] was proposed, showing that correspondences can be found by maximizing mutual information, although its performance is not better than other less complex cost functions.

Mutual information has been also largely used for medical image registration. In this field, [4] proposes to combine mutual and gradient information, showing an improvement with respect to classical mutual information based formulations. A multiresolution scheme for matching, which is only based on mutual information, has been presented in [5] showing its viability. An advantage of multiresolution schemes is the information suppression, which allow to analyze the structure of the images with different level of details. Thus, the information of the current level can be enriched by using the prior knowledge collected from previous levels in the hierarchy. A strategy such as the one mentioned above is presented in [6], being restricted to probability propagation. In the current work not only mutual information (I) but also gradient information (I_G) are propagated through two different scale-space representations. The first one is based on a scale-space stack while in the second a pyramidal representation is used for comparisons.

The main contributions of the proposed approach are as follow: (a) evaluate mutual information performance once gradient information is embedded; (b) increase the discriminative power by means of a classical pyramidal representation; and (c) show the improvements by propagating mutual and gradient information (I and I_G). The proposed approach is evaluated with a large number of templates; up to our knowledge previous works were, on the one hand, specially devoted to the registration problem; and one the other hand, they were validated on few samples. The paper is organized as follows. Section 2 presents the proposed approach together with basic formulations. Experimental results and comparisons are given in Section 3. Finally, conclusions are detailed in Section 4.

2. PROPOSED APPROACH

This section presents the key concepts used for the derivation of the proposed scheme. Firstly, a concise description of the problem from the information theory point of view is introduced. Next, all the essential topics that composes the proposed solution are covered, including: entropy, mutual information, scale-space representation, and the use of gradient information with mutual information. Finally, the proposed scheme for mutual information propagation, in two scale-space representations, is introduced.

2.1. Problem statement

The template matching problem could be stated as: let $\overline{I_l}$ be the neighborhood of a pixel with local image coordinates $I_l(x, y)$ (usually referred in the literature as *template window*); and a searching

This work has been supported by the projects TRA2007-62526/AUT and CTP-2008ITT00001 and research programme Consolider-Ingenio 2010: MIPRCV (CSD2007-00018).

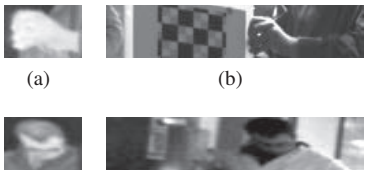


Fig. 1. Examples used for testing. *(left)* Infrared templates to be matched. *(right)* Corresponding search space (intensity images).

space consisting of a set of regions, referred as *matching windows* ($\overline{I_r}$) extracted from another image (I_r), the challenge is to find the best matching.

In the current work, without loss of generality, the matching windows $\overline{I_r}$ will move through the epipolar line that relates two rectified images. Therefore, the searching space is a collection of windows centered on $I_r(x+d, y)$ that does not belongs to the same modality of the template. The parameter d is the disparity, and represents the candidate disparity during the matching process. Figure 1(a) shows an infrared template $\overline{I_l}$ and Fig. 1(b) its searching space (intensity image). Another illustration is presented in Fig. 1(c) and Fig. 1(d).

2.2. Image template matching and information theory

An interesting point of view about information theory was presented in [7], which could be useful in the multimodal matching problem. Based on this theory, we reformulate the matching problem. The main change that has to be done, is to assume that both template and matching windows are two information sources, which produce a succession of symbols in a random manner. A more simple way to visualize this is to associate the symbols with image pixels, or other implicit information. Strictly speaking, let $(\Omega, \mathcal{B}, P)$ be a probability space, where Ω denotes a sample space, $\mathcal{B}$ a space of events defined as σ-field $\mathcal{B}$ of subsets of Ω, and a probability measure P that assigns a real number $P(F)$ to every member F of the σ-field $\mathcal{B}$. In our case, the probability space $(\Omega, \mathcal{B}, P)$ is the mathematical generalization of the interaction between: (a) symbols that could be intensity or infrared measurements; (b) the space of all possible output symbols, usually referred as alphabet; (c) events, which are sequences of symbols that can be drawn by sources of information; and (d) a probability measure assigned to events.

We can assume that $I_l(x, y) : \Omega \rightarrow A_{I_l}$ and $I_r(x+d, y) : \Omega \rightarrow A_{I_r}$ are two random variables defined on $(\Omega, \mathcal{B}, P)$ with alphabets A_{I_l} and A_{I_r}, respectively. There are several ways to define these alphabets, the most frequent, and used in this paper, consists in quantizing the pixel values and binning them in order to obtain a discrete alphabet $A_f = \{a_1, a_2, ..., a_{\|A_f\|}\}$.

2.2.1. Entropy

The entropy of a discrete alphabet A_f of random variables defined on the probability space (Ω, β, P), with a probability mass function p_f, is defined as:

$$H(A_f) = -\sum_{a \in A_f} p_f(a) \log p_f(a). \quad (1)$$

Usually, entropy is associated with a measure of randomness; it does not depend on the current values of pixels of $\overline{I_l}$ or $\overline{I_r}$, but on the probability distribution of A_f. So, a low entropy of A_f can be interpreted as a non random alphabet, where there is no uncertainty. For example, texture-less regions have a lower entropy than a highly textured region. In order to relate two different sources of information, without an apparent correspondence, it is necessary to find some content in common. As it was mentioned above, the link or bridge that perform this task is the joint entropy, concept related to mutual information.

2.2.2. Mutual information

Mutual information I is estimated by establishing alphabets to encode the pixels of $\overline{I_l}$ and $\overline{I_r}$. The mutual information is then obtained from every encoded pair. It is defined as:

$$I(\overline{I_l}; \overline{I_r}) = \sum_{a_i \in A_{I_l}} \sum_{b_j \in A_{I_r}} p_{I_l I_r}(a_i, b_j) \log \frac{p_{I_l I_r}(a_i, b_j)}{p_{I_l}(a_i)\, p_{I_r}(b_j)}, \quad (2)$$

where p_{I_l} and p_{I_r} are the marginal probability mass functions of alphabets and $p_{I_l I_r}$ the joint probability mass function. The alphabets A_{I_l} and A_{I_r} are built by normalizing each window independently (range [0, 1]) and then quantizing them into Q levels. The joint probability mass function $p_{I_l I_r}$ is a 2-dimensional matrix. Before being normalized, it holds the number of times symbol (a_i, b_j) is observed in $\overline{I_l}$ and $\overline{I_r}$. Notice that the alphabets are symbolized as $A_{I_l} = \{a_1, a_2, ..., a_N\}$ and $A_{I_r} = \{b_1, b_2, ..., b_N\}$. The marginal probabilities are determined by summing along each dimension of the previous matrix.

2.3. Scale-space representation

This section presents the basic notions about scale-space, which is used to build two data structures and to evaluate the scheme of propagation. Firstly, a scale-space stack representation is presented. Then, a pyramidal representation, which is faster than the previous one, is described.

2.3.1. Scale-space stack

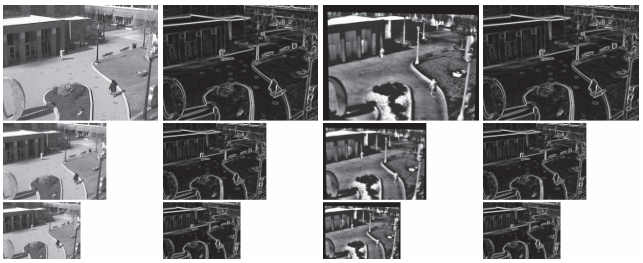
The scale-space representation $L : \mathbb{R}^N \times \mathbb{R}^+ \rightarrow \mathbb{R}$ for an arbitrary dimension N is obtained by convolving an image with a Gaussian derivative kernel of order n. Notice that, the zero scale is also included and corresponds to the given image. Following the notation presented in [8]:

$$L_n(\mathbf{x}; t) = g_n(\mathbf{x}; t) * I_k, \quad (3)$$

where $\mathbf{x} = (x_1, ..., x_N)^T \in \mathbb{R}^N$, $t \in \mathbb{R}^+$ is the current scale level, I_k is the current image, and $g_n(\mathbf{x}; t)$ is the Gaussian derivative kernel of order n. If $n = 0$ the Gaussian function is obtained, otherwise its corresponding derivative kernel. In this paper, since only gradient information is required, L_0 and L_1 are computed for I_l and I_r. It means a stack of Gaussian blurred images and their corresponding first order derivative images.

2.3.2. Pyramidal representation

Another way to generate a scale-space representation is by means of a pyramidal hierarchy, which is similar to the method described above. It consists in adding a new stage after the Gaussian filtering, which apply a downscale algorithm, sampling the output image at a constant rate. In this work, we have explored the use of an half octave Gaussian pyramid of zero and first order [9]. This representation has been chosen due to the reduction factor; hence it assure an optimal propagation of mutual information. Figure 2 shows two pyramidal representations of three levels; (a) and (b) correspond to



(a) $L_0(\mathbf{x};t)$ of I_r (b) $L_1(\mathbf{x};t)$ of I_r (c) $L_0(\mathbf{x};t)$ of I_l (d) $L_1(\mathbf{x};t)$ of I_l

Fig. 2. Three level pyramidal representations of two images; (a) and (b) intensity image; (c) and (d) infrared image.

an intensity image, while (c) and (d) to an infrared image. Note that in the two coarser levels the image features are still available, in spite of their small sizes. See [8, 9, 10] for a detailed description on pyramidal representations.

2.4. Incorporating gradient information

The gradient information is incorporated following the formulation presented in [4]. Gradient images are calculated as mentioned above, using the $L_1(\mathbf{x};t)$ scale-space representation (pyramid or scale-space stack). It is important to notice that intensity and infrared changes are not necessarily equals (nor in orientation neither in magnitude). However, since both images depict the same scene, corresponding gradient vectors could appear in both modalities and their phase difference be near to 0 or π (phase or counter-phase). Therefore, these vectors could be used to unveil possible matchings. Let $\mathbf{x}$ and $\mathbf{x}'$ be two corresponding points that belong to $\overline{I_l}$ and $\overline{I_r}$, respectively. Then, their phase difference is defined as:

$$\theta(\mathbf{x};\,\mathbf{x}';t) = \arccos\left(\frac{L_{1,l}(\mathbf{x};t)\cdot L_{1,r}(\mathbf{x}';t)}{|L_{1,l}(\mathbf{x};t)|\,|L_{1,r}(\mathbf{x}';t)|}\right), \qquad (4)$$

where $L_{1,l}(\mathbf{x};t)\cdot L_{1,r}(\mathbf{x}';t)$ is the dot product of their gradient values in x and y direction. The phase difference (eq. 4) is weighted by a function $w(\theta;t)$ that penalizes those gradient vectors that are not in phase or counter-phase:

$$w(\theta) = \frac{cos(2\,\theta)+1}{2}. \qquad (5)$$

Finally, the gradient information is incorporated into the mutual information formulation as in [4]:

$$I_G(\overline{I_l};\,\overline{I_r};t) = I(\overline{I_l};\,\overline{I_r};t)\cdot G(\overline{I_l};\,\overline{I_r};t), \qquad (6)$$

where $I(\overline{I_l};\,\overline{I_r};t)$ is the mutual information and $G(\overline{I_l};\,\overline{I_r};\,t)$ is the gradient information, computed as:

$$G = \sum_{\mathbf{x}\in\overline{I_l},\,\mathbf{x'}\in\overline{I_r}} w(\theta(\mathbf{x};\mathbf{x'};t))\,\min(|L_{1,l}(\mathbf{x};t)|\,|L_{1,r}(\mathbf{x}';t)|). \qquad (7)$$

2.5. Propagating mutual and gradient information

The current work proposes to improve the discriminative power of mutual information in two ways: (a) propagating *mutual information* through a scale-space stack and a pyramidal representation; and (b) propagating both *mutual and gradient* information through a scale-space stack and a pyramidal representation. Note that [6] proposes a similar scheme but for propagating joint probability ($p_{I_l I_r}$); while, our approach directly spread the I and I_G between consecutive levels, allowing changes in the sizes of the bins that represent the sources of information. This supposes a great advantage because at each level an optimum alphabet can be used, which is unsuitable in a scheme such as the one proposed by [6].

Our approach starts by computing I and I_G at a coarser level. Notice that in the case of a stack all the images in the stack have the same size. Thus, $I(\overline{I_l};\overline{I_r};t)$ and $I(\overline{I_l};\overline{I_r};t-1)$ have an interscale correspondence and the next equation (8) can be directly applied. However, in the case of a pyramidal representation the level $t-1$ contains a smaller image than the one in the current level t (downsampling). Therefore, two situations must be considered: (a) if $\mathbf{x}$ is not present in the previous level, only its value at the current level is considered (I or I_G) and the term λ in eq. (8) is set to 1; and (b) if $\mathbf{x}$ is present in the previous level, then a cubic spline interpolation is used to compute its I_{prior}, since we are using rectified images only ancestors on the epipolar line are considered (one dimensional interpolation problem); thus I_{prior} is obtained from its neighborhood at $(t-1)$. The propagation rule is defined as:

$$I_{current}(\mathbf{x};t) = \lambda\, I_{current}(\mathbf{x};t) + (1-\lambda) I_{prior}(\mathbf{x};t-1), \quad (8)$$

where λ is the confidence of current I or I_G.

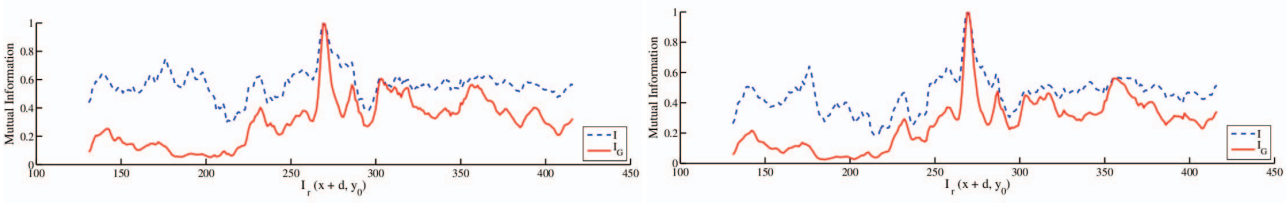
3. EXPERIMENTAL RESULTS

In order to evaluate the proposed approach, small parts of an infrared or color images are cropped and used as templates—61600 patterns in total were extracted from OTCBVS Benchmark Dataset [11]. I and I_G are cost functions computed between the template and all possible windows on the corresponding searching space; they are obtained without disparity restrictions. The correct match is located at point d where the cost function reaches the maximum value $\arg\max_d\{I(\overline{I_l(x,\,y)};\overline{I_r(x+d,\,y)})\}$, similarly for I_G.

The matching cost of a template and a candidate is obtained by computing I (eq. (2)) and I_G (eq. (6)). For example, Fig. 3(a) shows the results when the template depicted in Fig. 1(c) and the searching space in Fig. 1(d) are matched. Once the cost over the whole searching space is computed the three largest local maximum values are extracted (only three values were selected just for the sake of presentation simplicity). These values are used to quantify the results, which are depicted in Tables 1 and 2. Since color and infrared images in [11] are registered, the correct matches are known before hand. Then, it is possible to determine the correct one among the three local maximum selected above (a tolerance range of 2 pixels for d is used). Tables 1 and 2 present the percentages of correct matching that corresponds to first, second or third position. If a $winner-takes-all$ scheme was used, then the number of correct matches will be just the first column of these tables. The proposed approaches, both using a scale-space stack and a pyramidal representation, have been compared with the results obtained when I and I_G are not propagated through the different levels of the stack/pyramid (Tables 1 and 2 part (a)).

Figure 3(b) shows the results for the same example introduced above but when I and I_G are propagated. Note that both approaches (with/without propagation) find the correct match but by using propagation the relative values between local maximums are increased, making easier to identify the correct one.

Upper levels of scale-space stacks were obtained by convolving the images with a Gaussian kernel of order $n = \{0,\,1\}$ and $\sigma = \{1,\,2\}$, as shown the Table 1. The experiments were conducted following the next setup, in both the stack and the pyramid cases. The window size decreases with the scale. It started with a size of 32×32 and finishes with 8×8 (level 0); the propagation



(a) without propagation (b) with propagation

Fig. 3. (a) Mutual information (I) and mutual with gradient information (I_G) as formulated in [4]. (b) Proposed propagation of mutual information I and mutual with gradient information I_G (I: dashed line; I_G: solid line).

Table 1. First three maximums by using a scale-space stack

(a) Score without propagation (results in percentages)						
Level	1st.		2nd.		3rd.	
	I	I_G	I	I_G	I	I_G
2	37.19	50.03	12.52	11.55	7.08	5.83
1	17.30	27.58	9.61	9.95	7.10	5.11
0	**4.15**	**12.54**	**3.62**	**7.44**	**3.37**	**5.92**
(b) Score with propagation (results in percentages)						
2	37.19	50.03	12.51	12.52	7.08	5.83
1	31.69	49.61	12.57	12.57	7.71	5.96
0	**15.24**	**43.97**	**9.19**	**11.99**	**7.00**	**5.74**

Table 2. First three maximums by using a pyramidal representation

(a) Score without propagation (results in percentages)						
Level	1st.		2nd.		3rd.	
	I	I_G	I	I_G	I	I_G
2	31.29	54.04	15.86	1386	10.82	8.11
1	14.14	29.28	9.66	13.38	7.91	8.95
0	**4.15**	**12.54**	**3.62**	**7.44**	**3.37**	**5.92**
(b) Score with propagation (results in percentages)						
2	31.29	54.04	15.86	1386	10.82	8.11
1	17.67	39.55	15.31	15.53	7.10	8.36
0	**9.23**	**27.38**	**6.41**	**9.87**	**5.15**	**6.52**

also follows this direction. The parameter λ controls the degree of propagation between consecutive levels. Experiments have shown that $\lambda = 0.5$ maximizes the scores. The quantization parameter Q is constant ($Q = 30$).

I and I_G showed a behavior proportional to the size of template ($\overline{I_l}$). If it is increased then the estimation of I will be better, due to large number of observations. Nevertheless, big windows are not desirable for stereo matching. Therefore, our propagation scheme is a good choice because it improves the results whereas small windows (8×8 pixels) are used. The improvement obtained with the scale-space stack reaches about 3.5 times at the last level, while in the pyramidal representation it is about 2.2 times due to the downsampling.

The representations only have three levels in order to compare both results. The results of pyramid using propagation are better than without it. However, these results cannot be compared to the ones obtained with the stack, except at level 0, due to compression of images (see Fig. 2). Notice that, each level contains less information and the image is smaller; hence, the estimation of I is weak. The used mutual information estimator (eq. 2) and the way to ensemble the alphabets, establish a dependency between the estimation (I value) and the number of members in the sample (template size), which affects the performance of propagation in this representation.

Results presented in Table 1 and Table 2 show the improvements reached when gradient information is used with mutual information, instead of mutual information alone. On average, I_G improves the result from I about 3 times.

4. CONCLUSIONS

This paper presents a scheme for combining mutual information with gradient information together with an evaluation of two scale-space representations. Experimental results show the improvements in the discriminative power as well as the viability of the proposed approach. Future work will study a mutual information estimator robust to downsampling.

5. REFERENCES

[1] Paul Viola and William M. Wells III, "Alignment by maximization of mutual information," *IJCV*, vol. 24, no. 2, 1997.

[2] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imaging*, vol. 16, no. 2, pp. 187–198, April 1997.

[3] Geoffrey Egnal, "Mutual information as a stereo correspondence measure," Tech. Rep., University of Pennsylvania, 2000.

[4] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imaging*, vol. 19, no. 8, pp. 809–814, 2000.

[5] J.P.W. Pluim, J.B.A. Maintz, and M.A. Viergever, "Mutual information matching in multiresolution contexts," *Image and Vision Computing*, vol. 19, no. 1-2, pp. 45–52, 2001.

[6] Clinton Fookes, Anthony J. Maeder, Sridha Sridharan, and Jamie Cook, "Multi-spectral stereo image matching using mutual information," in *3DPVT*, 2004, pp. 961–968.

[7] Robert M. Gray, *Entropy and information theory*, Springer-Verlag, Inc., New York, USA, 2009.

[8] Tony Lindeberg, *Scale-Space Theory in Computer Vision*, Kluwer Academic Publishers, Norwell, MA, USA, 1994.

[9] James L. Crowley, Olivier Riff, and Justus H. Piater, "Fast computation of characteristic scale using a half-octave pyramid," in *In: Scale Space 03: 4th Inter. Conf. on Scale-Space theories in Computer Vision*, 2002.

[10] A. Kuijper, "Mutual information aspects of scale space images," *PR*, vol. 37, no. 12, pp. 2361–2373, 2004.

[11] James W. Davis and Vinay Sharma, "Background-subtraction using contour-based fusion of thermal and visible imagery," *CVIU.*, vol. 106, no. 2-3, pp. 162–182, 2007.