Visible-Light Image Synthesis from Infrared Images Using Texture Transfer

Yu Wakabayashi†, Sei Ikeda†, Kosuke Sato†

Abstract  We propose a novel method for synthesizing a visible-light image from a near-infrared one taken from the same viewpoint. Generating a multi-channel color image from a single-channel one is generally an ill-posed problem. As texture and color cues, the method refers to visible-light color images of the same scene but taken from a different point. The proposed method focuses on the estimation of the luminance component rather than the chrominance one because chrominance is not sensitive for human vision, and relatively easier to be estimated than luminance. Our cost function to optimize the luminance component is designed with the assumption that the synthesized image should be globally similar to the input infrared image structure and locally similar to the reference image patterns. Our experimental results show that the proposed method produces an artificial visible-light image, whose color appearance is more natural than color conversion methods, and geometrically more accurate than texture transfer-based methods.

Key words: near-infrared image, luminance component estimation, texture transfer, inpainting

1. Introduction

IR cameras are widely used for security, monitoring, earth observation, biometrics, intelligent transport systems (ITS) and so on. IR camera images are frequently represented by grayscale images to show observers. The absence of color may detract from a suitable comprehension of the scene because an object recognition depends on the stored knowledge of object’s chromatic characteristics. False color representation makes the reaction time for comprehension of the scene lesser, but an image colorized in a manner totally different from that of the actual scene continues to detract from the scene. Therefore, to synthesize a natural-looking visible-light images from IR images, some color images tend to be used as cues of realistic color appearance.

Many methods which synthesize an artificial visible-light image by referring real color images are proposed. These methods can be divided into multi-band (since 2003) and single-band approaches (since 2005). The former approach simply maps a set of pixel intensities of different IR bands to a single color based on its relationship to the reference image. Color transfer based methods adjust the gain and bias of the color image to be closer to these of the reference image. Color-table methods preliminary make color table from pixel-by-pixel correspondences IR images and the reference image or artificial visible-light image captured from the same viewpoint. Calculating cost is low in this approach, although it is impossible to assign different colors to pixels with the same IR intensity.

The latter approach calculates patch-by-patch correspondences considering some texture similarity such as the mean and standard deviation, the HOG feature, and learned feature. This approach enables more flexible colorization that can overcome the above mentioned disadvantage. However, this approach has a critical limitation that it is still difficult to reproduce realistic color appearances because IR intensity is assigned as the luminance component of the visible-light image even though both are greatly different as shown in Fig.1. Only Toet’s method estimates the luminance component by a linear transformation of an IR image. However, transformation using monotone function cannot resolve the differences between intensity of IR image and the luminance of visible-light one.

In this paper, we propose a novel method for synthesizing visible-light images from near-infrared images by estimating both chrominance and luminance components. As cues for color and texture, this method refers color images captured in the same scene but from different viewpoints. This condition assumes some fog-see-through application domain. In fog-see-through situations, there assumes to be fog or haze in front of a viewer and an IR camera at the same viewpoint. IR light images provides only geometric information from
the viewpoint because IR light can pass through fog or haze more than visible one. In this assumption, color information can be provided by visible-light color cameras as a part of ITS infrastructures on some public place such as highways, railways, and runways. In such environments, vehicles can obtain their own view images with IR cameras, and can receive slightly different view images as references, but similar to their own view, from other vehicles running in front or from cameras installed in the environments. Our method is potentially better than the method proposed by Fredembach et al.\textsuperscript{24)}, where the color camera must be placed at the same view point.

The proposed method focuses on the estimation of the luminance component rather than the chrominance one because once a good luminance component is obtained, the chrominance component can be easily computed from reference images by existing colorization methods. Our cost function to optimize the luminance component is designed with the assumption that the synthesized image should be globally similar to the input infrared image structure and locally similar to the reference patterns. For the chrominance, the method assigns color using the estimated correspondences.

2. Related research

The most relevant method in terms of dehazing applications is the method proposed by Fredembach et al.\textsuperscript{24)} This method synthesizes a dehazed visible-light image from both IR and color images taken from the same viewpoint. Our research tries to relax the viewpoint limitation to obtain more correct color information.

The proposed method can be regarded as a novel-view synthesis because the method produces an image of a different viewpoint from that of the reference image. Early approaches of novel-view synthesis have strongly relied on epipolar geometry and photo-consistency among multiple views; therefore, it has been difficult to achieve both geometric connectivity\textsuperscript{25)} and photo-realism\textsuperscript{26,27)}. One of the first successful work coping with this trade-off was done by Fitzgibbon et al.\textsuperscript{28)}. The cost function of their method is designed so that every pixel satisfies the photo-consistency and naturalness constraint which is commonly used in the patch-based texture generation approaches mentioned below. In our approach, instead of photo-consistency, the edge of the IR image is used to constrain patch search area\textsuperscript{29)}.

The proposed method is also similar to other patch-based image generation techniques such as texture transfer\textsuperscript{30,31)}, texture synthesis\textsuperscript{30,32)}, and inpainting\textsuperscript{29,33–35)}. Although the aims of these techniques are different, the common point is a constraint that similar patterns of local regions overlapping each other should be found in the reference one. Those methods limit candidates in patch search to form desired structures in the synthesized image. For example, the texture transfer method proposed by Efros et al.\textsuperscript{30)} searches the best patch only from the candidates which have similar patterns to the target local patch in the original image. The inpainting method using PatchMatch algorithm limits the patch search region on user-specified positions\textsuperscript{29)}. In the proposed method, the edge of the IR image is used to limit the patch search area.

3. Visible-light image synthesis

The proposed method consists of three processes. First, dense correspondences between the IR and reference images are roughly calculated in a preprocessing described in section 3.1. Second, the luminance component of the visible-light image is estimated using the dense correspondences as initial values. These dense correspondences are also optimized at the same
time. Finally, chrominance component is simply assigned from the reference image to the visible-light one using the optimized dense correspondences. In this section, we mainly describe the preprocessing and the luminance component estimation in the order.

3.1 Preprocessing

The purpose of the preprocessing is to calculate dense correspondences and to reduce rotational discrepancies between the IR and reference images because the patch-base image generation in the estimation of the luminance component requires good initial values of luminance and dense correspondences, and high computational cost to achieve rotational invariance. The preprocessing consists of two steps: computation of sparse correspondence and dense one.

Sparse correspondences between the IR and reference images can be obtained by a certain feature matching algorithm. Although it is difficult to obtain a large number of sparse correspondences because existing feature point descriptors do not work well for matching between IR and visible-light images, at least four sparse correspondences are required in order to estimate the homography referred to in the next step. If an IR image sequence or some color images are available as input, structure-from-motion (SFM) algorithm is available and its strong geometric constraint reduces incorrect correspondences. We employed the latter method in the experiment mentioned in section 4.

The homography can be directly calculated by sparse correspondences or indirectly obtained by each camera pose calculated by an SFM algorithm. The homography transforms the reference image to match the IR image, and reduces rotation errors. While dense correspondences calculated in this process are not accurate, they are used for initialization and are optimized in the estimation of the luminance component.

3.2 Estimation of luminance component

This process estimates the luminance component. The pixel-wise correspondences $f_{den}$ and the homography-transformed image obtained in the preprocessing are initial values for the optimization of the luminance component $Y_{vis}$ and the correspondence map $f_{pat}$ : $\Omega \rightarrow \Omega$, respectively. In the optimization, we minimize the following energy function $E$ combining the global similarity $E_{global}$ to the IR image and the local similarity $E_{local}$ to the reference image.

$$E(Y_{vis}, f_{pat}) = \sum_{x \in \Omega} \left\{ E_{global}(x, f_{pat}(x)) + E_{local}(x, f_{pat}(x)) \right\}, \quad (1)$$

where $\Omega \in \mathbb{R}^2$ is the image domain.

1) Local similarity to reference image

The purpose of the energy $E_{local}$ is to bring each neighboring pattern of point $x$ in the synthesized image close to the luminance pattern in the reference image. The most straightforward similarity measure is the sum of squared difference in a local window $T$ of the synthesized image $Y_{vis}$ and reference image $Y_{ref}$.

$$E_{local}(x, X) = \sum_{p \in T} (Y_{vis}(x+p) - Y_{ref}(X+p))^2 \quad (2)$$

2) Global similarity to IR image

The purpose of the energy $E_{global}$ is to reflect the IR image structure in the synthesized image. The energy $E_{global}$ is capable of limiting patch search area to only the pixels which have similar edges both in the IR and reference images. The edge similarity is evaluated in the sense of both edge strength $E_{str}$ and edge direction $E_{dir}$.

The energy $E_{global}$ is expressed by the multiplication of these two energies as shown in Eq.3.

The similarity error of the edge strength becomes small either if both edge strength $|\nabla Y_{IR}(x)|$ and $|\nabla Y_{ref}(x)|$ are higher than their threshold $T_{in}$ and $T_{ref}$, respectively, or if both are lower as shown in Eq.4. Otherwise, the error becomes infinite.

The similarity error of the edge directions becomes small when angle $\theta$ formed edges between the IR and reference images is smaller than the threshold $T_\theta$. The error becomes infinite only when each edge strength is higher than its threshold, in other words the edge direction is reliable. We employ the sine function to accept the reversal of luminance as shown in Fig.1.

$$E_{global}(x, X) = E_{str}(x, X)E_{dir}(x, X) \quad (3)$$

$$E_{str}(x, X) = \begin{cases} 1 & (|\nabla Y_{IR}(x)| < T_{IR} \text{ and } |\nabla Y_{ref}(x)| < T_{ref}) \\ 1 & (|\nabla Y_{IR}(x)| > T_{IR} \text{ and } |\nabla Y_{ref}(x)| > T_{ref}) \\ \infty & \text{otherwise} \end{cases} \quad (4)$$

$$E_{dir}(x, X) = \begin{cases} \sin^2 \theta & (|\nabla Y_{IR}(x)| > T_{IR} \text{ and } |\nabla Y_{ref}(x)| > T_{ref}) \\ \infty & \text{otherwise} \end{cases} \quad (5)$$

where gradient of each point $\nabla Y_{IR}$ and $\nabla Y_{ref}$ are calculated by the Sobel filter.

Since the edge existence should be judged independently of scene contrast, the thresholds $T_{in}$ and $T_{ref}$ are determined by the following equation.

$$T_{in} = \frac{Y_{max} - Y_{min}}{c}, \quad T_{ref} = \frac{Y_{max} - Y_{min}}{c} \quad (6)$$
where $Y_{\text{max}}$ and $Y_{\text{min}}$ are maximum and minimum luminance values in each window $W_{\text{pat}}$ of the IR and reference images, and $c$ is a constant value.

Fig. 2 show the effectiveness of the global similarity in comparison with two naive methods: The former is $E_{\text{global}} = E_{\text{str}}$ and the second is $E_{\text{global}} = E_{\text{dir}}$. In this comparison, we use the set of IR and reference images shown in Fig. 3(a) and (b), as shown in Section 4. The same parts of the estimated luminance component (a road marking) are shown in Fig. 2. The red dashed line in Fig. 2 indicates the correct position. Neither naive methods can not fix the positional gaps as shown in Fig. 2(a) and (b). In contrast, the proposed method can effectively adjust this gap as shown in Fig. 2(c).

(3) Energy optimization

The proposed method minimizes the energy $E$ (Eq.1) by using a greedy algorithm similar to Wexler’s EM approach. The following two processes are repeated until energy function $E$ converges: (i) Minimizing $E(f_{\text{pat}})$ with the luminance component $Y_{\text{vis}}$ fixed. (ii) Minimizing $E(Y_{\text{res}})$ with the corresponding map $f_{\text{pat}}$ fixed. The optimization finishes when the updates of $Y_{\text{res}}$ and $f_{\text{pat}}$ are completed. In the following section, we introduce a method to avoid local minima.

In order to avoid local minima efficiently and reduce the computational cost, a coarse-to-fine approach is employed. An image pyramid is generated and energy minimization processes, (i) and (ii), are repeated from higher-level to lower-level layers successively using a local window $T$. A result of higher-level is given to the lower layer as an initial value. At the lowest-level (original size), the energy optimization process is repeated while reducing in stages the local window $T$ in order to produce fine textures.

The reference image transformed by homography $H$ as the highest-level initial image. At the regions where the transformed reference image does not overlap, the average of the pixel value on the boundary of the homography-transformed image is given as an initial luminance value. The window $W_{\text{pat}}$ centered by the initial value of dense correspondences obtained in the preprocessing is set for the searching area for the corresponding point $f_{\text{pat}}(x)$.

4. Experiment

4.1 Image acquisition & comparing method

To demonstrate the validity of the proposed method mainly by qualitative analysis, we compared the proposed method with naive methods using two kinds of real scenes: outdoor and bookshelf. The outdoor scene contains many kinds of objects, i.e., both artificial and natural objects as shown in Fig. 3(i), and has different depths such as a road marking at about 4m and a center building at about 45m. The bookshelf scene shown in Fig. 3(ii) contains repeating patterns and various colors. The scene distance is approximately 2m, and there is an occluder, i.e., a center book.

For each scene, to obtain an input IR image and a color ground truth image taken from the same viewpoint, we used an IR light pass filter (720 - 2750 nm) or a visible light one (390 - 690 nm), respectively, which is mounted on a digital camera with a CMOS sensor sensitive to a band from 200 nm to 1100 nm. The captured IR and ground truth images are shown in Fig. 3(a) and (d), respectively. The reference images shown in Fig. 3(b) was taken with the same visible-light pass filter, but from the different viewpoint. For the outdoor and bookshelf scenes, the distances between the two viewpoints were approximately 40 cm and 20 cm, respectively.

As input of an SfM algorithm, Bundler was employed. For outdoor and bookshelf scenes, we also obtained nine IR and nine visible-light images captured from random viewpoints whose standard deviations were approximately 40 cm and 20 cm, respectively. The resolution of all the images was 640 × 480 pixels. The camera was preliminary calibrated by the OpenCV calibration tool. In the optimization, the image scale and the size of square windows $T$, $W_{\text{pat}}$ was set as shown
in Table 1 in accordance with the implementation by Kawai et al. 36). The constant c and threshold $T_\theta$ shown in the same table were empirically determined.

In this section, we compare visible-light images generated by three methods: Toet’s method 20), only local similarity, and the proposed method. Toet’s method estimates luminance component by the linear transformation of the IR image. We assigned a chrominance component of the ground truth image to that of the synthesized image by Toet’s method because this method does not assume such a complex scene. Only local similarity is a part of the proposed method considering only local similarity to the reference image ($E = \sum E_{local}$). The proposed method considers both the global similarity to the IR image and the local similarity to the reference image ($E = \sum E_{global} + E_{local}$).

4.2 Result

The estimated visible-light images are shown in Fig. 4 and 5. Label A to D indicate characteristic parts in each synthesized image. These results show the advantage of our method in the sense of the following three points.

The first point is the color appearance of synthesized images. The color appearance of the synthesized images by Toet’s method shown in Fig. 4(a) and Fig. 5(a) look totally different from the ground truth images shown in Fig. 3(d). In contrast, the visible-light images produced by the proposed method shown in Fig. 4(c) and Fig. 5(c) look similar in color to the ground truth.

The second is the magnitude correlation of luminance values. We can find this example at least at two parts: trees and their background in Fig. 4(a). The trees around Label B are brighter than their background. However, this relationship is inverse, and was effectively recovered by the luminance estimation as shown in Fig. 4(b) and (c).

The third is positional gap to the ground truth. The red dashed lines in Fig. 3, 4 and 5 are manually drawn and indicate the correct positions of each object. In Fig. 4, positional gaps can be found in the image of (b) especially at the road marking (Label A) and buildings (Label B and C), but in (c), large gaps can be found relatively few. As a negative effect, a potted plant (Label D) found in Fig. 4. does not appear in (c).

In the results of the bookshelf scene in Fig. 5, similar effects can be found. Large positional gaps can be found in (b) at the lower right part of Label B. Although a red book is partly occluded by a book indicated by Label A in (b), the corresponding book is synthesized in (c). However, the texture of the left part around Label C in (c) is more blurred than that in (b).

In addition to the qualitative analysis, the peak signal-to-noise ratio (PSNR) and the structural similarity (SSIM) are shown at the bottom of each image in Fig. 4 and Fig. 5. The proposed method has the near best score of the three.

5. Discussion

Although we can simply obtain a visible-light image
apparently similar to the ground truth by homography transform of the reference image, this transformed image still has some positional gaps. The proposed method reduced these gaps at least around Label A, B, and C in both scene images by using geometric constraints of the domain for searching corresponding texture. However, there are some undesired effects such as the potted plant (Label D) shown in Fig. 4(c). In these parts, edge of each object is not well extracted as shown in the rightmost column of Fig. 6(Label D). This result shows that edge extraction is sensitive to the capability of fixing positional gaps and transferring good texture. We leave room for further studies on what edge extraction operator for this purpose is the best.

In addition, there is another limitation concerned with the homography-transformed image used as an ini-
tial value of the luminance component. The optimization of this method highly depends on the initial value. Therefore, at least, if the transformed image has larger gaps than the radius of the window for searching similar patches, or if local rotations or scale difference cannot be absorbed by the homography transformation as shown around Label C in Fig. 5, the proposed method cannot find good texture.

It is also basically difficult to estimate the luminance component at the regions where the reference image does not overlap because we cannot set the initial value or find good textures from the reference. For estimating the luminance in these regions, we can also use additional multiple reference images if they are available or can be acquired from the Internet.

In our experiment, an SfM algorithm was employed to obtain the corresponding feature points satisfying epipolar constraints between the IR and reference images. This is still paradoxical even though a large number of correspondences are not required for estimation of homography as mentioned in Section 3.1. A most hopeful method is a registration between two sets of 3D points estimated by an SfM because 3D structure is independent of wavelength.

In the qualitative analysis, the PSNR and SSIM were employed as criteria for evaluation. PSNR and SSIM levels are approximately 20-40 dB and 0.8-1.0, respectively, if the synthesis success. Both values shown in Fig. 4 and 5 indicate that the proposed method does not have the capability of reproducing the pixel-level structures, even though the appearance looks better. This characteristic is similar to other texture synthesis techniques.

6. Conclusion

In this paper, we proposed a novel method for synthesizing an artificial visible-light image from a near infrared image using a visible-light image taken from a different viewpoint. In contrast to existing methods, the proposed method firstly focused on estimating the luminance component. The luminance component estimation follows an optimization of the local continuity and global registration, which is a kind of inpainting algorithms with a geometric restriction. The experimental results show the feasibility of the proposed method at least under the condition that disparity of corresponding pixels between infrared and visible-light images is not large. In two examples, we found that the color appearance of the synthesized images was more natural than a previous color conversion method, and geometrically more accurate than an inpainting method.

7. Acknowledgments

This work is partially supported by JSPS KAKENHI Grant Number 25730122, 15H02737 and 15K12084.

References


