RGB-D DATA FUSION IN COMPLEX SPACE

Ziyun Cai

Electronic and Electrical Engineering
University of Sheffield
Mappin Street, Sheffield S1 3JD, U.K

Ling Shao

School of Computing Sciences
University of East Anglia
Norwich NR4 7TJ, U.K

ABSTRACT

Most of the RGB-D fusion methods extract features from RGB data and depth data separately and then simply concatenate them or encode these two kinds of features. Such frameworks cannot explore the correlation between the RGB pixels and their corresponding depth pixels. Motivated by the physical concept that range data correspond to the phase change and color information corresponds to the intensity, we first project raw RGB-D data into a complex space and then jointly extract features from the fused RGB-D images. Consequently, the correlated and individual parts of the RGB-D information in the new feature space are well combined. Experimental results of SIFT and fused images trained CNNs on two RGB-D datasets show that our proposed RGB-D fusion method can achieve competing performance against the classical fusion methods.

Index Terms— RGB-D fusion, complex space

1. INTRODUCTION

Single RGB image understanding has been studied very well over the past decades. Nevertheless, many challenges still exist in computer vision research area because of the limited information provided by RGB images. With the invention of high-quality and low-cost depth sensor (*e.g.* the Microsoft Kinect), a new stream of research turns to seek new types of image representations for overcoming the traditional hard tasks [1] [2] [3]. Socher *et al.* [4] present a model which is based on the combination of Convolutional Neural Networks (CNNs) and Recursive Neural Networks (RNNs), but this model extracts features of RGB-D images separately. Song *et al.* [5] propose an approach which makes a $3D$ volumetric scene from RGB-D images as input and $3D$ object bounding box as output through Region Proposal Network to learn objectness and a joint $2D$+$3D$ object recognition network to extract geometric features in $3D$ and color features in $2D$.

However, above mentioned methods have not explored the correlation between raw RGB images and raw depth images. Most of the methods just learn features from RGB and depth separately and then simply concatenate them together as RGB-D features or encode these two kinds of features. The major disadvantage is that the correlation and complementary property between RGB and depth are ignored, and learning procedures cannot be adjusted mutually. To better explore the correlation between the RGB pixels and the corresponding depth pixels, and take advantage of the complementary property, we first project raw RGB-D data into a complex space and then jointly learn features from the fused RGB-D images. The correlated and individual parts of the RGB-D information in the new feature space are well combined. Our fusion method can also be considered as representing the data closer to the nature of the data. In physics, the range data correspond to the phase change and color information corresponds to the intensity. From computer vision view, the feature representations are expected to satisfy low mutual information and also show a lot of variations. The fused RGB-D data should be treated holistically. We modify the classical SIFT to evaluate our fusion method. Moreover, experimental results on AlexNet [6] also show that our fusion method is advantageous. It is worthy to note that modified SIFT is just an example to show the advantages of the fusion method. Other RGB-D methods can also be introduced into complex space. The main contribution of this paper is a new method which can better explore the correlation between the RGB pixels and the corresponding depth pixels for RGB-D data fusion. It makes the correlated and individual parts of the RGB-D information in the new feature space well combined.

2. FUSION METHODOLOGY

The fused images in our methodology are represented by complex values, which are closer to the nature of the data itself. This methodology makes the representations of RGB-D images greater distinctiveness, higher entropy on the whole images, higher entropy of the scale-space derivatives and larger feature quantity. RGB images can be considered as amplitude measurements, which depend on the nature illumination. According to the depth images, no matter which sensing system is chosen for depth image representations, the pixel values of the depth images always mean the distances from the camera to the observed objects. The depth image is often considered as the phase change measurement, which

depends on the measured scattering received from the active illumination with the sensor. The phase can be regarded as actual distance. For depth measurements with the uniqueness range, the well-known inverse-square law of active intensity reveals the approximation: $I_D \approx I_R/\phi^2$, where I_R is the intensity of the RGB image, I_D is the intensity of the corresponding depth image, ϕ is the phase which can be calculated from the depth values d. Therefore, it proves that it is reasonable to correlate I_R and I_D together through (I_R, I_D, ϕ). In physics, the phase difference can always be considered as a phase value from the mathematical concept, hence the representations of the RGB-D images with complex values become natural. Note that all the RGB images mentioned in our methodology are first converted into gray images. We define $I_R(x, y)$ as the RGB image, $I_D(x, y, d)$ as the depth image, where $d = d(x, y)$, x and y are the image coordinate points, d is the depth value on the coordinate (x, y). Combining the physical and mathematical concepts, the fused complex-valued image function is expressed as:

$$f(x, y, d) = I_R(x, y) + I_D(x, y, d)e^{i\phi(x,y,d)}, \tag{1}$$

where $\phi = \phi(x, y, d) \in [0, 2\pi)$ is defined through the range:

$$d = n \cdot 2\pi\ell + \phi\ell, \tag{2}$$

where $n \in \mathbb{N}$, $2\pi\ell$ is the uniqueness range of the camera with $\ell \in \mathbb{N}$. The natural number ℓ is a multiple of some unit of length, and n is the "wrapping number". Moreover, with the representation of complex number, the fused image can be represented as Polar representation or Cartesian representation:

$$I_f^P = |f|e^{iarg(f)},\ I_f^C = Re(f) + iIm(f), \tag{3}$$

where in above equation, $|f| = \sqrt{I_R^2 + 2I_R I_D \cos\phi + I_D^2}$, $arg(f) = \arctan\frac{I_D \sin\phi}{I_R + I_D \cos\phi}$, $Re(f) = I_R + I_D \cos\phi$ and $Im(f) = I_D \sin\phi$. Since we normalize all complex-valued images, we can obtain $max|f| = 1$. Fig. 1 shows some example images from different scenes. Since Kinect cannot perceive the light-reflecting area, we discard the scenes which include many mirrors and tiles, such as bathroom. Each row of Fig. 1 includes RGB image, gray image, depth image, fused image and 3D graphic simulation of one scene. From the images produced by our fusion function in the fourth column, we can visually find that the fused images do contain the depth information and hold the color data simultaneously.

Till now, we have obtained all image representations in this paper: $I_R(x, y)$ for RGB images, $I_D(x, y, d)$ for depth images, I_f^P for the fused RGB-D Polar representations and I_f^C for the fused RGB-D Cartesian representations. In the following subsections, we will give the comparison among these representations from three aspects.

2.1. Mutual Information and Independence

If the mutual information among the images is low, it means that these images have more independence. According to [7],

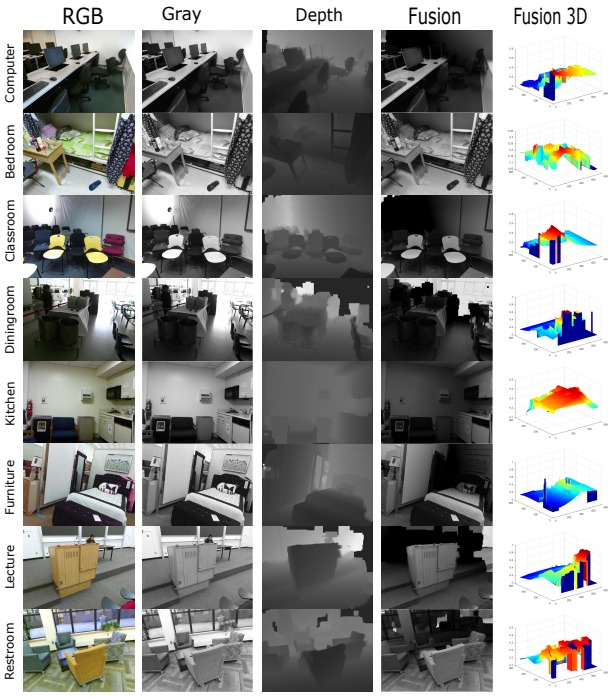

Fig. 1: Some random example images from 8 different scenes.

good features should be distinctive. The information content depends on the representations of the images [8]. Therefore, we expect that the fused RGB-D images have the property of distinctiveness while containing more entropy. We can easily obtain that $E(I_f^P)$ or $E(I_f^C) > E(I_R)$ or $E(I_D)$. Moreover, information theory [8] describes that the image representation transformed from Polar coordinate to Cartesian coordinate increases entropy. Meanwhile, the procedure of our fusion method adds structure information into the original gray images, which increases the entropy. Therefore, based on above conclusions, we can obtain that $E(I_f^C) > E(I_f^P) > E(I_R)$ or $E(I_D)$. In the following discussions, we only compare higher entropy image presentations: $E(I_f^C)$ and $E(I_f^P)$.

We define $E_{A,\omega}$ as the joint entropy of $A = |f|$ and $\omega = arg(f)$ and $E_{R,I}$ as the joint entropy of $R = Re(f)$ and $I = Im(f)$. We can obtain:

$$E_{A,\omega} = E_{R,I} + \langle logA \rangle, \tag{4}$$

where from differential entropy representation,

$$\langle logA \rangle = \int \rho(R, I) logA(R, I) dRdI, \tag{5}$$

where $\rho(R, I)$ is the joint distribution function of R and I. Through the Jacobian transformation between the distributions, we can obtain $\rho(R, I) = \rho(A, \omega) \cdot |J|$. Under this circumstance, since $max|f| = 1$, $J = A(R, I) = \sqrt{R^2 + I^2} \leq 1$. Meanwhile, it proves $\langle logA \rangle < 0$.

The mutual information of (R, I) and (A, ω) can be defined as:

$$MI(R, I) = E_R + E_I - E_{R,I}, \tag{6}$$

$$MI(A,\omega) = E_A + E_\omega - E_{A,\omega}. \tag{7}$$

Therefore, with Eq. (4), the difference between $MI(R,I)$ and $MI(A,\omega)$ is as following:

$$\begin{aligned} \nu :&= MI(R,I) - MI(A,\omega) \\ &= (E_R + E_I) - (E_A + E_\omega) + \langle logA \rangle. \end{aligned} \tag{8}$$

Here, ν is the measure for mutual information and independence between $MI(R,I)$ and $MI(A,\omega)$. Since the smaller $MI(x,y)$ is, the more independent x and y are, if $\nu < 0$, it means $MI(R,I)$ is more independent than $MI(A,\omega)$. From the information theory, the value of $(E_R + E_I) - (E_A + E_\omega)$ is really small, which hardly affects the value ν. Meanwhile, we have proved $\langle logA \rangle < 0$. Actually, in our experiments, both ν and $\langle logA \rangle$ are computed around -1.

2.2. Feature Distribution

Better representation is more uniformly distributed on the image plane. With the increase of the entropy from different image representations, the number of the extracted features with the same method increases as well. Since we have proved that $E(I_f^C) > E(I_f^P) > E(I_R)$ or $E(I_D)$, I_f^C is supposed to have the most features among above four representations.

Take image scale-space feature detection for example, the image representation from Polar to Cartesian increases the entropy of the scale-space derivatives. With the increase of image derivative entropy, it leads to more persistent texture. The scale-space equation $\frac{\partial f}{\partial t} = \triangle f$ aims to find the persistent texture. Since I_f^C contains more entropy than I_f^P over the scale-space derivatives, we define:

$$E_{\dot{A},\dot{\omega}} = E_{\dot{R},\dot{I}} + \langle A \cdot |cos\dot{\omega}sin\omega - sin\dot{\omega}cos\omega| \rangle, \tag{9}$$

where $\dot{A}$, $\dot{\omega}$, $\dot{R}$ and $\dot{I}$ are the derivation of f on A, ω, R and I. In this case, $\langle A \cdot |cos\dot{\omega}sin\omega - sin\dot{\omega}cos\omega| \rangle < 0$.

Since $E(I_f^C) > E(I_f^P) > E(I_R)$ or $E(I_D)$, it follows from the Jacobian $J = A \cdot (cos\dot{\omega}sin\omega - sin\dot{\omega}cos\omega)$ of the transformation of derivatives. The RGB features and the depth features in real-valued image pairs are defined to be a scale-space feature for I_R and I_D. Similarly, for the scale-space feature, a Cartesian feature and a Polar feature in a complex-valued image are defined for $Re(f)\&Im(f)$ and $|f|\&arg(f)$ respectively. Therefore, we can obtain the comparison of the number of features about these four representations: $N(I_f^C) > N(I_f^P) > N(I_R)$ or $N(I_D)$.

2.3. Euclidean KS-distance to Uniformity

Robust features are always sampled from a uniform distribution. If the features are closer to the uniform, it means that the features contain more entropy. We consider the extracted n features from a scene are independent from another, and these features are identically distributed. According to n independent and $\{X_1, X_2, \cdots, X_{i1}\} \in \mathbb{R}$ as identically distributed random variables with common cumulative distribution function $F(x)$, empirical distribution function $F_n(X)$ is defined:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mu_{(-\infty,x]}(X_i), \tag{10}$$

where $\mu_{(-\infty,x]}$ is the indicator function. The extracted n features mentioned above can be considered as be taken from $F(x)$. From Glivenko-Cantelli theorem, F_n uniformly converges to F:

$$\|F_n - F\|_\infty = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \longrightarrow 0. \tag{11}$$

For arbitrary F, $\|F_n - F\|_\infty$ is called as the Kolmogorov-Smirnov statistic [9]. Since it has properties of distance between cumulative distribution functions, it is also called as KS-distance. We can define different image representation KS-distances through Euclidean norm:

$$d(S,\lambda) := \sqrt{\|F_n(x) - \lambda_x\|^2 + \|F_n(y) - \lambda_y\|^2}, \tag{12}$$

where S can be considered as n points in the plane. λ_i and λ are the cumulative distribution functions of the uniform distribution on the coordinate axis and on the plane separately. $d(S,\lambda)$ is called as Euclidean KS-distance to uniformity. We define S_C as the sample of Cartesian features, S_P as the sample of Polar features, S_R as the sample of RGB features and S_D as the sample of depth features. According to the independence, $S_C > S_P > S_R$ or S_D, and the number of features is $N(S_C) > N(S_P) > N(S_R)$ or $N(S_D)$, we can obtain:

$$d(S_C,\lambda) < d(S_P,\lambda) < d(S_R,\lambda) \; or \; d(S_D,\lambda), \tag{13}$$

where λ is the uniform distribution on the image plane.

Therefore, in our RGB-D fusion methodology, I_f^C is chosen as the optimal image representation.

3. EXPERIMENTAL SETUP

We evaluate our work on the NYU Depth V1 dataset [10] and SUN RGB-D dataset [11]. The classical SIFT is designed for real-valued images, which makes it no sense on complex-valued images. Therefore, we modify SIFT from essences. Different from SIFT which detects the local extrema and minima of $D(\sigma)$ through comparing its 26 real-valued neighbors, our algorithm chooses to compare the module m among these neighbors. The module can be calculated as $m^2 = Re(f)^2 + Im(f)^2$. It can make sure that the color information and the depth information are all considered when choosing the keypoints. Examples of two pairs of RGB-D images by modified SIFT are shown in Fig. 2. From Fig. 2, we can see that the keypoints detected in the fused images are much more than the sum of keypoints in raw RGB images and keypoints in raw depth images under the same parameters of keypoint detection. It shows the advantages of our fusion method. At last, HOG $3D$ [12] is chosen in our algorithm to

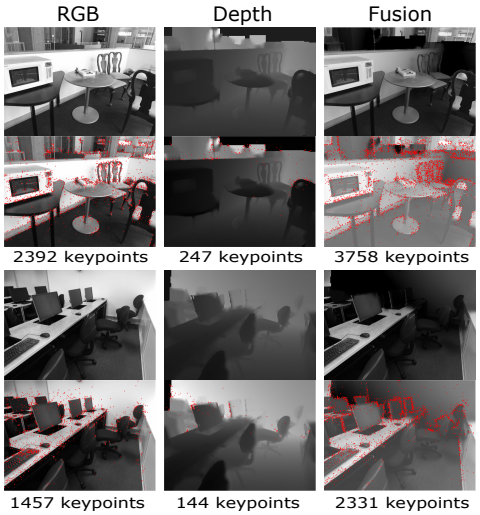

Fig. 2: Examples of two pairs of RGB-D images by modified SIFT. The parameters of keypoint detection are same.

Table 1: Accuracies (%) for scene classification on NYU Depth V1 and SUN RGB-D datasets.

Methods	NYY Depth V1		SUN RGB-D	
	SIFT	AlexNet	SIFT	AlexNet
RGB	55.0	60.1	19.2	22.3
Depth	50.3	54.2	17.7	20.6
RGB-D	59.6	65.7	21.3	28.7
Our method	**63.4**	**70.1**	**24.4**	**31.2**

describe the key points. We choose depth as the third dimension. For details on HOG $3D$, see [12].

In addition, for deep features, we use the NYU depth V2 RGB-D dataset [13] with more than 200K frames from the 249 training video scenes for learning the fused images initial AlexNet [6]. Then we fine-tune the fused images from NYU Depth V1 and SUN RGB-D on this initial model, to extract the features of the fc-7 layer and train SVM on top of it. The feature dimension after CNNs is 4096. Note that for traditional RGB-D fusion method, the depth image is encoded as HHA image as in [14] before extracting the features. Then the features extracted from both RGB and depth are concatenated together as inputs for SVM classifier. In this article, modified SIFT and fused images trained CNNs are examples to show the advantages of the fusion method. It inherits the principles of our fusion method and fully takes advantages of depth information through being extended to a complex space. The model for classification is linear SVM.

3.1. Experimental Results

The baselines in NYU Depth V1 and SUN RGB-D datasets are calculated through SIFT extracted from RGB images and SIFT on the depth images. The RGB-D SIFT features are created by concatenating RGB features and depth features together. Our RGB-D modified SIFT features and fused images CNNs features are from the images produced by our fusion method on the whole dataset.

Table. 1 shows that the classification accuracy of our fusion method outperforms RGBD-SIFT by around 4% and 3.1% respectively in NYU Depth V1 and SUN RGB-D datasets. Moreover, according to the fused images trained AlexNet, with the same number of pre-trained images, our method outperforms RGB-D features by around 4.5% and 2.5% respectively. Note that since the AlexNet in [11] is pre-trained through over $2M$ scene images from Places dataset [15], our final results are less than the baseline in [11].

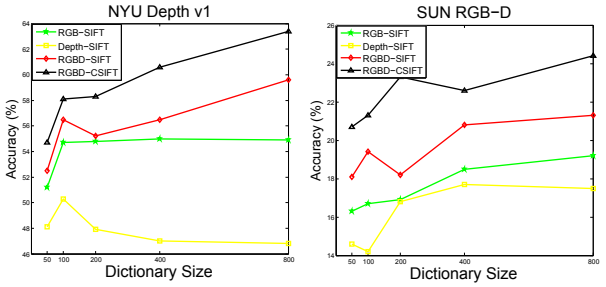

Fig. 3: Scene classification performance on NYU Depth v1 and SUN RGB-D datasets with different dictionary sizes.

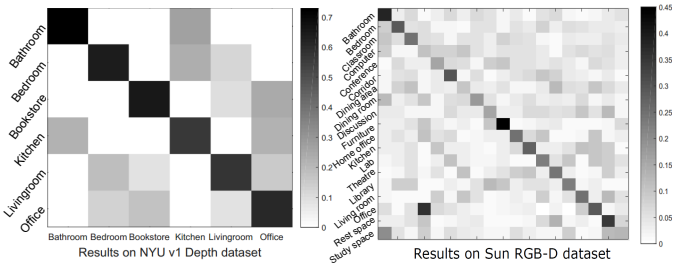

Fig. 4: Confusion matrices about our fusion method results on NYU Depth V1 and SUN RGB-D datasets.

In addition, for a better understanding of the comparison of our fusion method plus modified SIFT and the traditional SIFT, we choose different K-means dictionary sizes on two RGB-D datasets. Fig. 3 shows that a significant performance gain is obtained with a large dictionary. The confusion matrices across these two datasets are shown in Fig. 4.

4. CONCLUSION

In this paper, we have proposed a new RGB-D fusion method for fusing RGB-D images, which can better reveal the correlation between the RGB pixels and the depth pixels, taking advantage of the complementary property. The experimental results show that our method achieves competing performance against the classical fusion methods.

5. REFERENCES

[1] Ziyun Cai, Jungong Han, Li Liu, and Ling Shao, "Rgb-d datasets using microsoft kinect or similar sensors: a survey," *Multimedia Tools and Applications*, pp. 1–43, 2016.

[2] Ling Shao, Ziyun Cai, Li Liu, and Ke Lu, "Performance evaluation of deep feature learning for rgb-d image/video classification," *Information Sciences*, vol. 385, pp. 266–283, 2017.

[3] Mengyang Yu, Li Liu, and Ling Shao, "Structure-preserving binary representations for rgb-d action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1651–1664, 2016.

[4] Richard Socher, Brody Huval, Bharath Bath, Christopher D Manning, and Andrew Y Ng, "Convolutional-recursive deep learning for 3d object classification," in *NIPS*, 2012, pp. 665–673.

[5] Shuran Song and Jianxiong Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," 2016.

[6] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[7] Tinne Tuytelaars and Krystian Mikolajczyk, "Local invariant feature detectors: a survey," *Foundations and Trends® in Computer Graphics and Vision*, vol. 3, no. 3, pp. 177–280, 2008.

[8] David JC MacKay, *Information theory, inference and learning algorithms*, 2003.

[9] Wayne W Daniel et al., "Applied nonparametric statistics," 1990.

[10] Nathan Silberman and Rob Fergus, "Indoor scene segmentation using a structured light sensor," in *IEEE ICCV Workshops*, 2011, pp. 601–608.

[11] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.

[12] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008, pp. 275–1.

[13] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012, pp. 746–760.

[14] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*, pp. 345–360. 2014.

[15] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014, pp. 487–495.