

ADSP Final Project:

Image Similarity

電機碩一

r06921048

李友岐

Abstract:

Given two images, one is the original image and the other is a distorted version of first one. The researchers want to quantify the visibility of differences between an image and its distorted version. In unit 9, the professor introduced several methods to measure signal similarity, such as mean square error (MSE) and normalized root mean square error (NRMSE). Mathematically, The theory of NRMSE is rational. When we compute NRMSE of two images, however, it doesn't match the result that human perceive. Therefore, Zhou Wang proposed a new method to measure similarity between two images, which is called structure similarity (SSIM). [1] This method provides a result close to the perception of human. Although SSIM is very effective, it isn't perfect. If the image is produced by non-structural distortion, such as rotation and displacement, we can't recognize the origin image and distorted image by SSIM. As a result, Zhou Wang came up with another method called complex wavelet SSIM (CW-SSIM). [2] By this method, we can overcome the weakness of SSIM.

Index Terms:

Error sensitivity, human visual system (HVS), mean square error (MSE), Peak signal-to-noise ratio (PSNR), normalized root mean square error (NRMSE), perceptual quality, structural similarity (SSIM), complex wavelet SSIM (CW-SSIM).

I. Introduction:

Most of the digital images may face a wide variety of distortions during processing, compression, transmission and reproduction, any of which would lead to deterioration of image quality. If we want to quantifying visual image quality, the only way to guarantee the correctness is through subjective evaluation, which is biased, opinionated, and highly influenced by the person's feelings. In real-word engineering, subjective evaluation is normally time-consuming and expensive. As a result, it is

very inconvenient and impractical. Therefore, objective evaluation is certainly necessary. In order to achieve object evaluation, we need to develop a quantitative method to predict the quality of perceived image automatically. Objective evaluation plays a important roles in several image processing applications. First, we can use it to adjust image quality dynamically. Second, we can use it to optimize the parameter settings of image processing systems. Third, we can use it to benchmark image processing systems. We take original image as reference and compare it with distorted image. This approach is named as full-reference. It means that we already know the complete reference image. In real-word applications, however, there is no way the complete reference image is available. Therefore, it is hard to do that practically. In this paper, for the convenience of comparing results, we only focus on full-reference image quality assessment. [1]

The most widely used full-reference quality metric is the mean squared error (MSE), because its manipulation is extremely simple. we only need to compute the average of the squared intensity differences of reference image and distorted image. A slightly revised version of MSE is called normalized root mean square error (NRMSE). After normalization, we can guarantee the computed results vary from a smaller range, which is more appropriate to use. Peak signal-to-noise ratio (PSNR) is also a popular approach due to its convenience for calculation. These three methods can be easily optimized mathematically. In practice, however, they can't match the perceived visual quality. The computed results are highly different with the perception of human being. [1]

In the last 50 years, the researchers tried their best to develop the methods taking advantage of the characteristics of the human visual system (HVS). Most of these methods are modified through MSE measure. As a result, we can penalize errors (differences) accord with their visibility. In 2004, Zhou Wang proposed a new approach based on the hypothesis that the HVS is highly sensitive to structural information. This approach is called structural similarity (SSIM), it not only considers image deterioration as perceived change in structural information, but also incorporate luminance masking and contrast masking terms. [1] Normally, the pixels have high dependencies on neighbors. These dependencies provide important structure information of the object in visual scene. There are two tendencies of

image distortions. The first one is that distortions would be less visible in bright regions, while the second one is that distortions become less visible if a significant activity appears in the image. We call these two phenomena luminance masking and contrast masking, respectively. SSIM is accord with the visibility of human being, however, it still has some drawbacks. If the distortion is non-structural, such as rotation and displacement, SSIM would be less effective. Therefore, in 2005, Zhou Wang proposed another method called complex wavelet SSIM (CW-SSIM). [2] CW-SSIM extends the original SSIM approach to the complex wavelet transform domain and make it insensitive to these non-structural image distortions. By this method, we can overcome the weakness of SSIM.

Section II summarizes the formula and meaning of MSE, NRMSE and PSNR. Section III describes the definition of SSIM and compares the results of SSIM with NRMSE. In Section IV, we show the formula of CW-SSIM and compares the results of CW-SSIM with SSIM. Section V shows the experiment results, where we compare NRMSE, SSIM and CW-SSIM of different pairs of original image and distorted image. Section VI and Section VII are our conclusion and the references, respectively.

II. MSE, NRMSE and PSNR:

$$\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |y[m,n] - x[m,n]|^2$$

Fig. 1: MSE

$$\sqrt{\frac{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |y[m,n] - x[m,n]|^2}{\sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |x[m,n]|^2}}$$

Fig. 2: NRMSE

$$10 \log_{10} \left(\frac{X_{Max}^2}{\frac{1}{MN} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} |y[m,n] - x[m,n]|^2} \right)$$

X_{Max} : the maximal possible value of $x[m, n]$
In image processing, $X_{Max} = 255$

Fig. 3: PSNR

Fig. 1, Fig. 2 and Fig. 3 shows the formulas of MSE, NRMSE and PSNR, respectively. [3] MSE measures the average of the squares of the errors. In other words, the average squared difference between the reference image and distorted image. NRMSE is a normalized version of the measure of root mean squared difference between the reference image and distorted image. After normalization, the computed result is more convenient to compare with other results. PSNR measures the ratio between the maximum possible value of an image signal and the value of corrupting noise. The value of corrupting noise can be calculated by MSE. Since there are many signals having a extremely wide dynamic range, PSNR is expressed in terms of the logarithmic decibel scale normally.

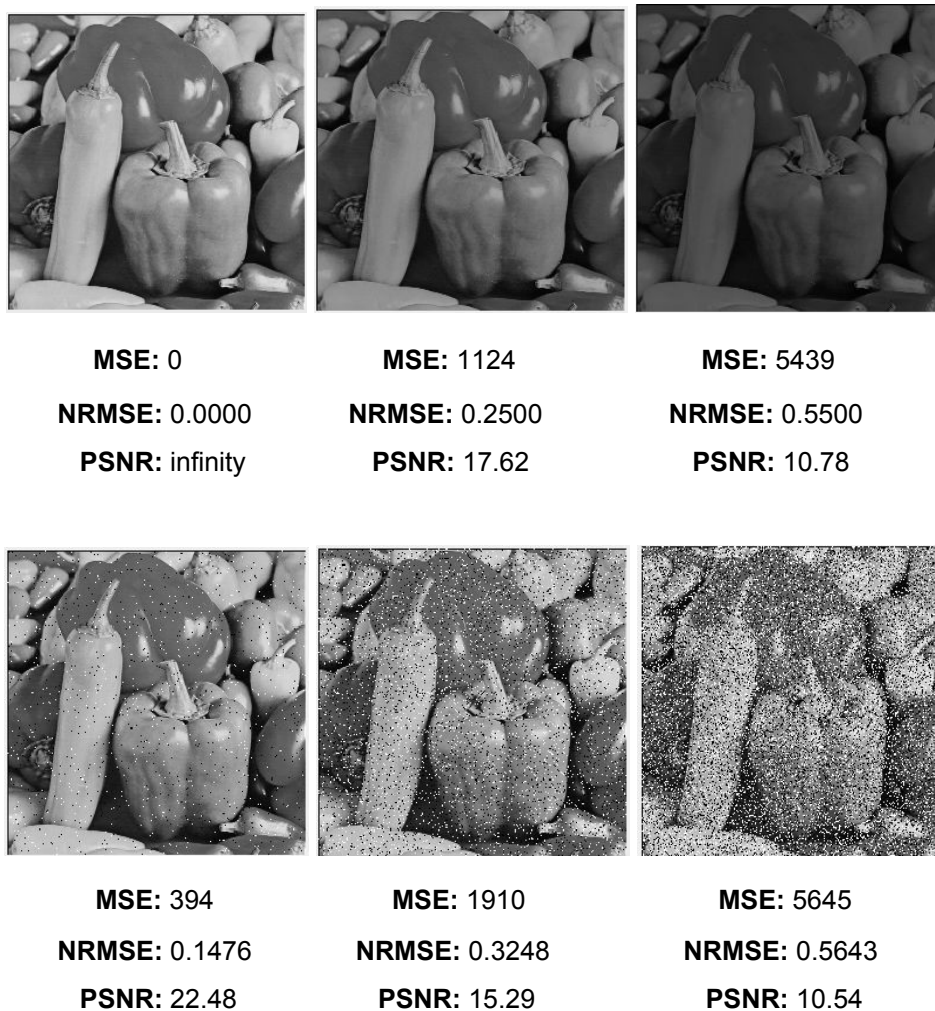


Fig. 4: An example of comparing images with MSE, NRMSE and PSNR.

Fig. 4 shows an instance of comparing reference image and distorted images with MSE, NRMSE and PSNR. There are six images in Fig. 4, where the first one is the reference image (Peppers.png) and the others are distorted images. The second one and the third one are the darker version of the reference image, while the last three images are created by adding pepper noise to the reference image. Note that if we compare two equivalent images, then MSE, NRMSE and PSNR of these two images would be zero, zero and infinity, respectively. According to the computed results, the larger the difference between reference and distorted image we perceive, the larger the value of MSE and NRMSE would be. In contrast, the value of PSNR decreases as the difference between reference and distorted image increases.

III. SSIM:

The pixels of a natural image have strong dependencies of each others, since the signal is highly structured. These dependencies carry important structure information about the visual objects. Despite the fact that most quality measures based on error sensitivity use linear transformations to decompose image signals, the strong dependencies between pixels are not removed. The motivation of this approach is to compare the structures of the original (reference) and the distorted images by a more straightforward method. [1]

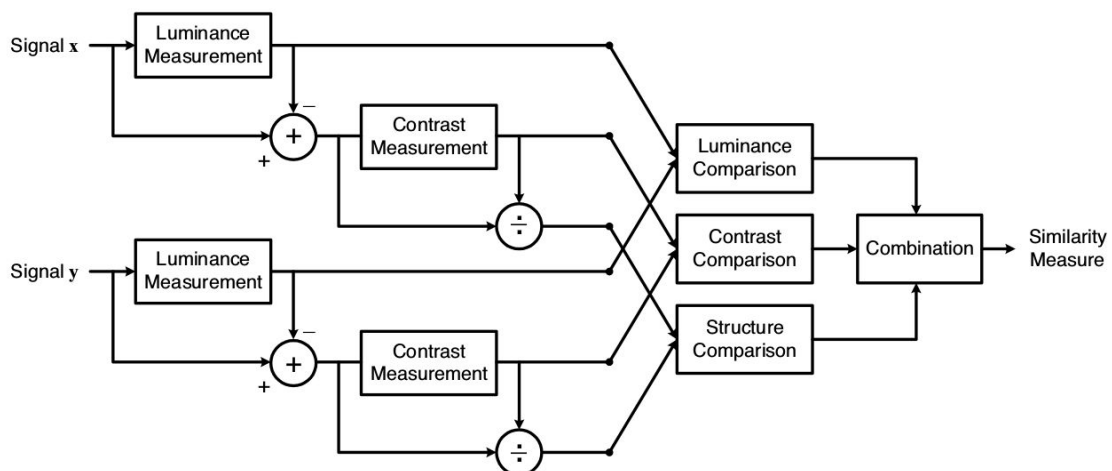


Fig. 5: Diagram of the structural similarity (SSIM) measure system

The product of the illumination and the reflectance represents the luminance of the surface of an object, but the illumination is independent of the structures of the objects. As a result, we need to separate the influence of the illumination in order to obtain the structural information of the image. The structural information of an image represents the structure of visual objects, which are independent of the luminance and contrast. We consider the local luminance and contrast only due to the fact that luminance and contrast could change under different scene. Fig. 5 is the diagram of proposed system. Given two nonnegative image signals, which have been aligned with each other. If we regard one image as perfect quality, then the quantitative measurement of the quality of the other image can be defined as the similarity between these two images. This similarity measure system consists of three comparisons, which are luminance, contrast and structure. [1]

$$S(\mathbf{x}, \mathbf{y}) = f(l(\mathbf{x}, \mathbf{y}), c(\mathbf{x}, \mathbf{y}), s(\mathbf{x}, \mathbf{y}))$$

Above is the overview of the proposed formula. $S(x, y)$ is the measurement of similarity. $l(x, y)$, $c(x, y)$ and $s(x, y)$ are the comparison of luminance, contrast and structure, respectively.

A. luminance

The luminance comparison function is define as

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

where μ_x and μ_y are the mean of x and y , respectively. Besides, C_1 is an adjustable constant.

B. contrast

The contrast comparison function takes a similar form

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}$$

where σ_x and σ_y are the standard deviation of x and y , respectively. Besides, C_2 is an adjustable constant.

C. structure

The comparison of structure is conducted after luminance subtraction and variance normalization. To measure the quantity of the structural similarity, we can apply the correlation between $(x - \mu_x) / \sigma_x$ and $(y - \mu_y) / \sigma_y$, which is effective and simple to calculate. Besides, the correlation between $(x - \mu_x) / \sigma_x$ and $(y - \mu_y) / \sigma_y$ is equivalent to the correlation coefficient between x and y . [1] Hence, the structure comparison function is defined as

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

where C_3 is an adjustable constant.

D. combine three components

Finally, we combine three functions $l(x, y)$, $c(x, y)$ and $s(x, y)$. The resulting similarity measure is named as the SSIM index between two images x and y . [1]

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^\alpha \cdot [c(\mathbf{x}, \mathbf{y})]^\beta \cdot [s(\mathbf{x}, \mathbf{y})]^\gamma$$

where α , β and γ are adjustable parameters. The relative importance of the three components are affected by the value of three parameters. Normally, we set α , β and γ all three parameters to 1. Besides, we set $C_3 = C_2/2$. [1] After setting these values, the expression can be simplified as

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Usually, The larger the difference (error) between reference and distorted image, the larger the value NRMSE is. The researchers found that, however, the result doesn't match our perception. Fig. 6 is an example showing the weakness of NRMSE. There are three images, where the first, second and third one are the reference image (Peppers.png), the darker version of reference image and an image with different structure (House.png), respectively. From our vision, the second one is more similar to the reference image than the third one. However, the third image has a smaller NRMSE result, which means it is more similar to the first image mathematically. Since this result isn't accord with what we perceive, we apply SSIM to compare these images. The larger the value of SSIM is, the more the two images are similar. Note that if we compare two equivalent images with SSIM, then the resulting value is 1.00, which means they are 100% similar. According to SSIM, the second image is 61% similar to the first one while the third one is only 17% similar to the first one. Obviously, SSIM method provides a more rational result.

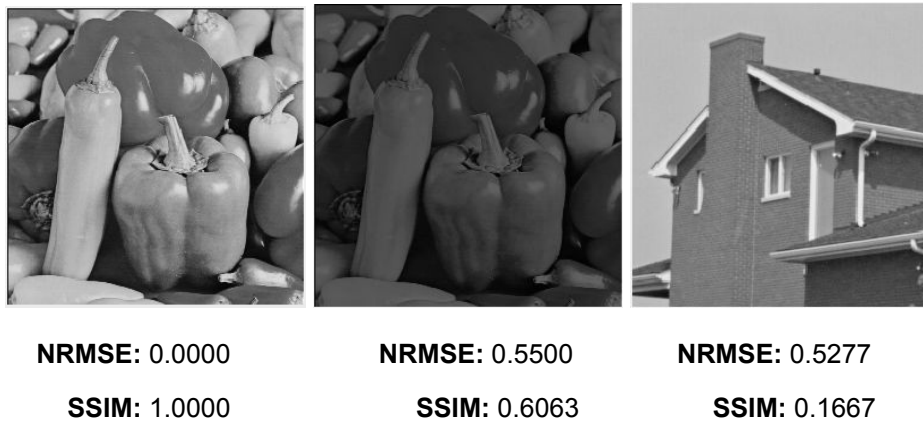


Fig. 6: An example of comparing images with NRMSE and SSIM.

IV. CW-SSIM:

We compare two images x and y through the complex wavelet transform domain. Let $c_x = \{c_{x,i} \mid i = 1, \dots, N\}$ and $c_y = \{c_{y,i} \mid i = 1, \dots, N\}$ denote two sets of

coefficients which are extracted at the same location of same wavelet. The SSIM approach can be extended to a complex wavelet SSIM (CW-SSIM) index as follow.

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2 \sum_{i=1}^N |c_{x,i}| |c_{y,i}| + K}{\sum_{i=1}^N |c_{x,i}|^2 + \sum_{i=1}^N |c_{y,i}|^2 + K} \cdot \frac{2 |\sum_{i=1}^N c_{x,i} c_{y,i}^*| + K}{2 \sum_{i=1}^N |c_{x,i} c_{y,i}^*| + K}$$

where c^* denotes the complex conjugate of c and K is an adjustable constant.

Because of the characteristic of the wavelet filters, the coefficients are zero mean.

The formula can be viewed as a product of two component. The first component is only affected by the magnitudes of the coefficients. If $|c_{x,i}| = |c_{y,i}|$ for all i 's, then the value of the first component become 1. In contrast, the second component is completely determined by the consistency of phase differences between c_x and c_y . If the phase difference between $c_{x,i}$ and $c_{y,i}$ for all i 's are same, then the value of second component become 1. Since the consistent phase shift of all coefficients of two sets doesn't affect the structure of image and the relative phase patterns of the wavelet coefficients contain most of the structural information of local image features, the second component is regarded as an effective measurement of the structural similarity of images. In addition, the numerator of the first component and the denominator of the second component are the same. [2] Consequently, The function can be further simplified as follow

$$\tilde{S}(\mathbf{c}_x, \mathbf{c}_y) = \frac{2 |\sum_{i=1}^N c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^N |c_{x,i}|^2 + \sum_{i=1}^N |c_{y,i}|^2 + K}$$

Although SSIM matches the perception of human being, there are still some special cases that SSIM isn't effective. Fig. 7 is a typical example of the weakness of SSIM. There are six images in Fig. 7. The first one is the reference image (Peppers.png). The second and third image are the darker version of the first one and another image with different structure (House.png). The fourth, fifth and sixth image are the reference image counterclockwisely rotating 90° , 180° and 270° ,

respectively. We can find that SSIM approach is too sensitive to rotating. The value of SSIM significantly reduces after rotation, which remaining 10 to 17% only. Moreover, the third image even has a larger SSIM than these three images created by rotating, which is extremely inconsistent with the visibility of human being. Since Peppers.png and House.png are two completely different images. In order to fix this error, we apply the extended approach CW-SSIM. It is obvious that the image created by rotating the reference 180° has a way better similarity, which is 71% CW-SSIM. The other two rotating images, however, still has smaller similarity than House.png. Therefore, we find that CW-SSIM is not perfect. It still has room for improvement. Rotating 90° and 270° should not lead to the low similarity as Fig. 7 shows.

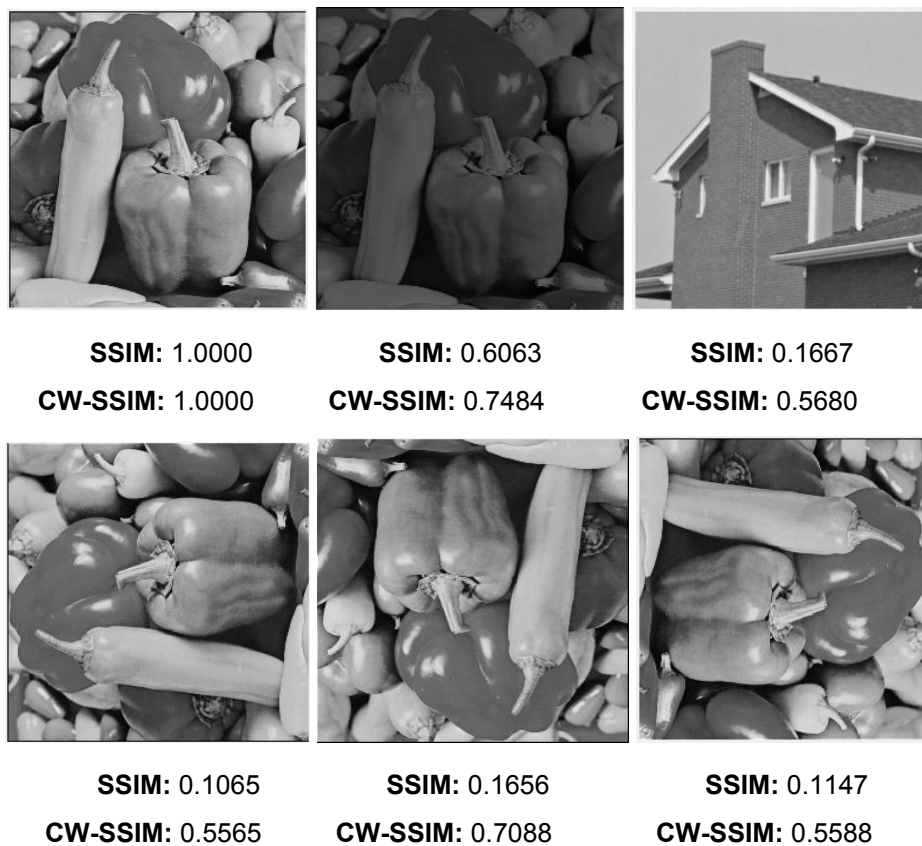
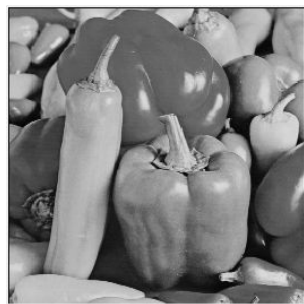







Fig. 7: An example of comparing images with SSIM and CW-SSIM.

V. Experiment results:

Fig. 8 shows an example that we compare NRMSE, SSIM and CW-SSIM of several images. The first one is the reference and the 2nd to 9th images are described in above sections. The 10th, 11th and 12th image are created by applying dilation, erosion and high pass filter on the reference image, respectively. It is obvious that CW-SSIM fits our visibility the best among these three approaches. Surprisingly, the images with pepper noise all maintain high CW-SSIM, exceeding 90%. This phenomenon indicates that pepper noise has a small influence on the value of CW-SSIM. In addition, dilation and erosion also slightly affect the value of CW-SSIM, remaining about 90%. CW-SSIM becomes, however, extremely low after using high pass filter. In the 12th image, there are only edges left, but the structure is still similar to the reference. Thus, I think the similarity between the reference and the 12th image should be higher. It should at least higher than the similarity between the reference and House.png, which would match with human perception more.

		
NRMSE: 0.0000	NRMSE: 0.5500	NRMSE: 0.5277
SSIM: 1.0000	SSIM: 0.6063	SSIM: 0.1667
CW-SSIM: 1.0000	CW-SSIM: 0.7484	CW-SSIM: 0.5680
		
NRMSE: 0.1476	NRMSE: 0.3248	NRMSE: 0.5643
SSIM: 0.9316	SSIM: 0.7154	SSIM: 0.4101
CW-SSIM: 0.9997	CW-SSIM: 0.9926	CW-SSIM: 0.9451

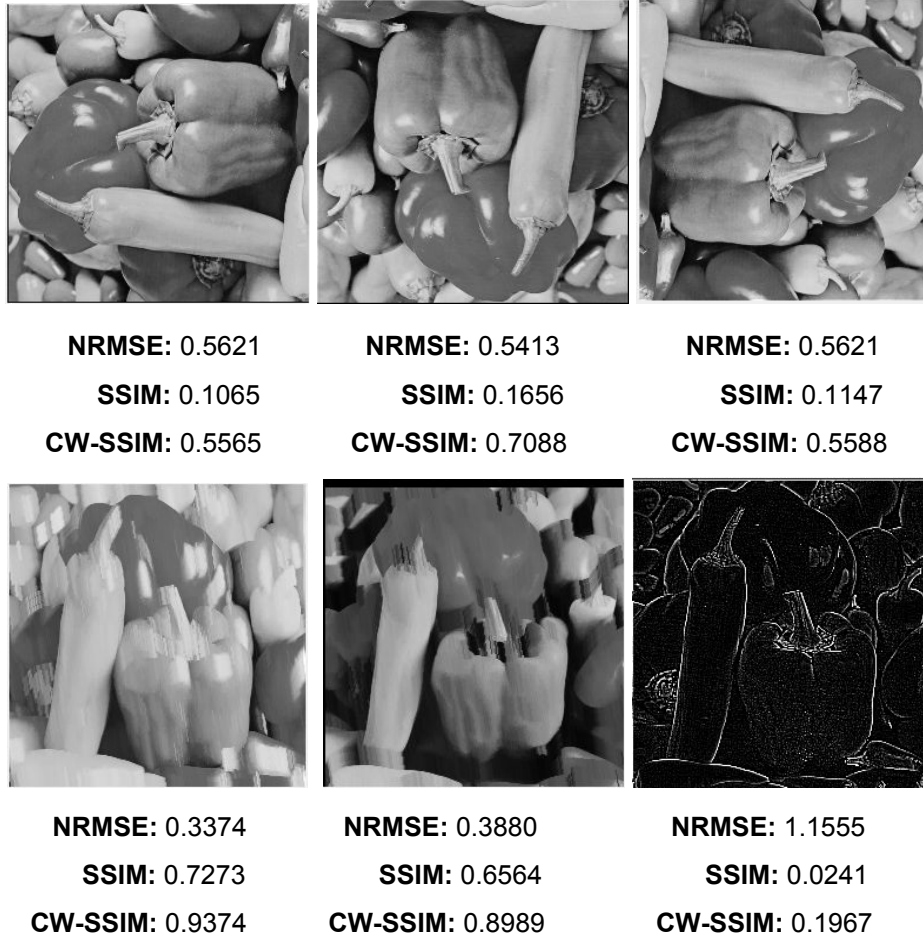


Fig. 8: An example of comparing images with NRMSE, SSIM and CW-SSIM.

VI. Conclusion and future work:

In this paper, we introduce several method to measure the similarity between two images, such as MSE, NRMSE, PSNR, SSIM and CW-SSIM. The approaches of MSE, NRMSE and PSNR are rational mathematically, but the result has a big difference with human perception. SSIM overcomes the drawbacks of the first three method. Nevertheless, it can't handle the distortion like rotation and displacement. As a result, CW-SSIM was proposed, providing a more rational result. However, it still has room for improvement. When we say those two images are similar, we normally mean the structures of them are similar. Hence, I think if we want to measure the similarity, we should increase the influence of structure information. The luminance should only have little impact on the image similarity. Besides, rotating

90°, 180° and 270° should result in almost same similarity. Figuring out a good enough solution to deal with these situation can be our future work.

VII. Reference:

- [1] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, pp. 600– 612, Apr. 2004.
- [2] Z. Wang, E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain", *Proc. IEEE Int. Conf. Acoustics Speech Signal Processing*, pp. 573-576, Mar. 2005.
- [3] Jian-Jiun Ding, "Lecture note Unit 9 of Advanced Digital Signal Processing " .