DEVELOPING A TRANSFORMER-BASED APPROACH FOR FUSING INFRARED AND VISIBLE IMAGES FOR IMPROVED OBJECT DETECTION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

AYTEKIN ERDOGAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF INFORMATION SYSTEMS

AUGUST 2023

Developing A Transformer-Based Approach for Fusing Infrared and Visible Images for Improved Object Detection

submitted by **AYTEKIN ERDOGAN** in partial fulfillment of the requirements for the degree of **Master of Science in Information Systems Department, Middle East Technical University** by,

Prof. Dr. Director of Institute _____
Dean, **Graduate School of Informatics**

Prof. Dr. Head of Department _____
Head of Department, **Information Systems**

Assoc. Prof. Dr. Supervisor _____
Supervisor, **Department, School**

Assoc. Prof. Dr. Co-supervisor if Exists _____
Co-supervisor, **Department, School**

Examining Committee Members:

Prof. Dr. Committee Member 1
Department, School _____

Assoc. Prof. Dr. Committee Member 2
Department, School _____

Assoc. Prof. Dr. Committee Member 3
Department, School _____

Assist. Prof. Dr. Committee Member 4
Department, School _____

Assist. Prof. Dr. Committee Member 5
Department, School _____

Date: 28.08.2019

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Aytekin Erdogan

Signature :

ABSTRACT

DEVELOPING A TRANSFORMER-BASED APPROACH FOR FUSING INFRARED AND VISIBLE IMAGES FOR IMPROVED OBJECT DETECTION

Erdogan, Aytekin

M.S., Department of Information Systems

Supervisor: Assoc. Prof. Dr. Supervisor

Co-Supervisor: Assoc. Prof. Dr. Co-supervisor if Exists

August 2023, **??** pages

English abstract here

Keywords: A keyword, another keyword, some other keywords

ÖZ

TÜRKÇE BAŞLIK

Erdogan, Aytekin
Yüksek Lisans, Bilişim Sistemleri Bölümü
Tez Yöneticisi: Doç. Dr. Supervisor
Ortak Tez Yöneticisi: Doç. Dr. Co-supervisor if Exists

Ağustos 2023, **??** sayfa

Türkçe öz buraya

Anahtar Kelimeler: Bir anahtar kelime, başka bir anahtar kelime, başka anahtar kelimeler

To the memories of my beloved friends Murat Tekin and Ragip Enes Katran

ACKNOWLEDGMENTS

Acknowledgments here

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

IF	Image Fusion
VIF	Visual and Infrared Image Fusion
AI	Artificial Intelligence
CNN	Convolutional Neural Networks
GAN	General Adveserial Networks
SR	Sparse Representation
MST	Multi-Scale Transformation
LRR	Low-Rank Representation

CHAPTER 1

INTRODUCTION

Image Fusion is a computer vision task that has been taken place for many years. Gathering all the complementary usefull informations into single image is called image fusion, *a.k.a* IF. Visual and Infrared Image Fusion, *henceforth will be referred to as VIF*, is a subfield of iamge fusion. Since the first study [**?**] in 1989, VIF is actively studied. In the era of AI, new methods such as CNN, GAN, auto-encoder, transformers are also applied to the VIF problem.

1.1 Research Questions

1.2 Contributions of the Study

1.3 Organization of the Thesis

CHAPTER 2

RELATED WORK

Image fusion algorithms can be categorized based on several factors, including whether they employ learning methods or hand-crafted steps, based on predefined loss functions, and whether labeled datasets are involved in the process. Learning-based methods involve the use of machine learning techniques, such as CNN, GAN, Transformers, Auto-encoders, to learn the features and relationships between input images. In contrast, hand-crafted approaches involve the manual selection and design of specific features and fusion rules.Whether a methods is end-to-end is defined as do they require one or more handcrafted steps. Loss functions are commonly used in learning-based methods to measure the quality of the fused image and to guide the training process. Based on the loss function, the process can be classified as self-supervised, supervised or unsupervised. Finally, labeled datasets can be used to train learning-based methods, where the some form of label is available and ground-truth labels are known, or for evaluating the performance of fusion algorithms. By considering these factors, researchers can select the most appropriate image fusion algorithm for their specific application requirements.

The available VIF methods can be categorized into two groups: traditional methods, which were widely used before the advent of AI, and learning-based methods. Regardless of the classification, all methods consist of three main components: image feature extraction, fusion of features from multiple images, and reconstruction of the image from the fused features. During the feature extraction stage, features are extracted from multiple images. In the fusion stage, the extracted features are compared, and complementary features are incorporated into a single feature map or set. Finally, in the reconstruction stage, the image is reconstructed from the fused set of features. All related studies aim to improve one or more of these stages in the process.

Learning-based fusion methods are often used to combine information from multiple sources to obtain more accurate and informative results. These methods can be categorized based on whether they require ground truth labels or not. The ground truth annotations don't have to be a form of fused image. If ground truth labels are needed, they are known as supervised methods. On the other hand, if no ground truth labels are required, they are referred to as unsupervised or self-supervised methods. These fusion techniques are commonly employed in convolutional-based methods, which are widely used in image processing and computer vision applications. Regardless of the specific method used, the goal of learning-based fusion methods is to improve the quality and accuracy of the resulting output by leveraging information from multiple sources.

Learning-based image fusion algorithms are commonly categorized based on the type of network used in the algorithm. There are several types of networks that are frequently used, including CNN-based,

auto-encoder based, GAN-based, transformer-based networks and others which include hybrid or handcrafted steps. Convolutional neural network (CNN)-based methods are often used due to their ability to extract features from input images and produce high-quality results. Auto-encoder-based methods are also popular, as they can effectively compress and decompress image information to obtain a fused image. Generative adversarial network (GAN)-based methods use a combination of generator and discriminator networks to produce high-quality fused images. Transformer-based methods have recently gained popularity due to their ability to process long sequences of input data efficiently. Other methods, such as sparse representation-based methods and wavelet-based methods, are also used in learning-based image fusion. Ultimately, the choice of network depends on the specific requirements of the task at hand and the available resources.

In addition to categorizing fusion algorithms based on the type of network used, another way to differentiate them is whether they are end-to-end. An end-to-end algorithm is one that can take raw input data and produce the desired output directly, without any intermediate hand-crafted steps. In the context of image fusion, end-to-end algorithms are those that can take multiple input images and produce a fused image without the need for manual feature extraction or other preprocessing steps. These algorithms are often preferred, as they can be more efficient and less prone to errors compared to non-end-to-end methods that require manual intervention. In contrast, non-end-to-end algorithms may require additional processing steps, such as registration and feature extraction, to produce the final fused image. While these steps can provide additional control over the fusion process, they can also introduce additional complexity and reduce the overall efficiency of the algorithm.

2.1 Traditional Algorithms

The traditional image fusion algorithms have been extensively studied in the literature. However, they are not without shortcomings. One of the major issues with these methods is the presence of handcrafted steps, which can lead to suboptimal results. In addition, the time complexity of some methods can be quite high, making them impractical for real-world applications.

Sparse representation (SR) based methods are a popular choice for image fusion. However, they suffer from several limitations. For example, methods such as [**?**] and [**?**] require dictionary learning, which can significantly increase the time complexity of the algorithm. Furthermore, these methods include handcrafted steps, which can limit their generalizability.

Another commonly used approach for image fusion is multi-scale transformation (MST) based methods. These methods, such as [**?**] and [**?**], can be quite effective at capturing various characteristics of images at different scales. However, they too suffer from limitations. One major issue is their lack of generalizability, which can make them less effective in certain scenarios.

Low-rank representation (LRR) based methods are another popular choice for image fusion. These methods, such as [**?**]liu2012robust, are particularly effective at dealing with noise and other forms of image degradation. However, like the other methods, they too have limitations. For example, they may not be suitable for all types of images, particularly those with complex textures or patterns.

In summary, the success of traditional image fusion algorithms heavily relies on the quality of the feature extraction method used. While there are many different methods available, each with its own

strengths and weaknesses, it is important to carefully consider the specific requirements of the problem at hand before selecting a particular method.

2.2 CNN Based Deep Learning Algorithms

CHAPTER 3

USER EXPERIMENT

In this chapter, the details of the user experiment are presented.

3.1 Research Method and Experiment Design

CHAPTER 4

USER EXPERIMENT

In this chapter, the details of the user experiment are presented.

4.1 Research Method and Experiment Design

CHAPTER 5

CONCLUSION AND FUTURE WORK

REFERENCES

[1] A. Toet, L. J. Van Ruyven, and J. M. Valeton, "Merging thermal and visual images by a contrast pyramid," *Optical engineering*, vol. 28, no. 7, pp. 789–792, 1989.

[2] Y. Bin, Y. Chao, and H. Guoyu, "Efficient image fusion with approximate sparse representation," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 14, no. 04, p. 1650024, 2016.

[3] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Optical Engineering*, vol. 52, no. 5, pp. 057006–057006, 2013.

[4] H.-M. Hu, J. Wu, B. Li, Q. Guo, and J. Zheng, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2706–2719, 2017.

[5] K. He, D. Zhou, X. Zhang, R. Nie, Q. Wang, and X. Jin, "Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain," *Journal of Applied Remote Sensing*, vol. 11, no. 1, pp. 015011–015011, 2017.

APPENDIX A

TABLES FOR RELATED WORK CHAPTER

A.1 Summary of the Studies

APPENDIX B

EXTRA MATERIAL

APPENDIX C

INSTRUMENTS AND ETHICAL CLEARANCE