学位報告４

報告番号	※甲　　　　第　　　　号

主論文の要旨

論文題目　Subjective Video Attribute Recognition: Labeling, Training and Explaining（映像における主観的属性の認識：ラベリング、学習、および解釈）

氏　　名　　吉　颖

論文内容の要旨

The analysis of video attribute recognition can be divided into two classes: Objective Attribute Recognition (OAR) and Subjective Attribute Recognition (SAR). The OAR is primarily concerned with recognizing tangible and stable characteristics, such as actions, objects, or scenes in a video. However, SAR is significantly more challenging since the attributes are defined by human perceptual cognition and are influenced by individual experiences or emotions. Subjective attributes such as image quality, aesthetics, and video popularity, are even difficult to define and recognize. The objective of this thesis is to investigate ground-truth generation, attribute recognition, and model explanations for subjective attributes. With this target, this thesis proposes two subjective video datasets, two training approaches, and one explainable module for SAR.

This thesis contains six chapters.

Chapter 1 describes the background and motivation of the work. The overview of the proposed methods is also concluded.

Chapter 2 introduces the related studies about OAR and SAR. Specifically, research relevant to ground-truth generation, subjective attribute recognition, and explainable artificial intelligence are presented.

To enhance the performance of SAR, this thesis addresses two practical challenges. Chapter 3 introduces an improved dataset labeling method designed for video violence recognition. Chapter 4 provides a training approach for improving the recognition accuracy of social relation atmosphere. Chapter 5 offers an improved model understanding method for explaining the inner procedures of 3D CNNs. By focusing on these three key difficulties of SAR, the recognition can be finally enhanced.

Chapter 3 introduces the process of constructing a clean dataset with reliable and stable ground truth. Currently, data annotations are often provided as single labels with majority voting. The annotators give an absolute value for the objective attribute. However, the subjective data do not have an exact ground truth. This chapter solves the problem by introducing a pairwise comparison method. The pairwise comparison can reduce the ambiguity and divergence in the annotating process. This chapter takes violence extent analysis as an example and provides a new dataset with subjective violence extent labels. Consequently, a rank learning method is specially designed to estimate the violence extent. Considering the violence extent of each video is a relative attribute and can be compared, the proposed method can learn the relationship between videos at the same level and different levels.

Chapter 4 introduces an efficient way to recognize and represent subjective attributes. End-to-end network is a commonly used method to recognize videos. However, the features directly extracted from the networks are not specific enough to represent the data. A new dataset is first created with both subjective social relation atmosphere attribute and objective social relationships attribute. A 3D explanation module proposed in Chapter 5 is used as a plug-in module in this chapter. The module can be used to extract the most important regions for recognizing social relationships. A framework is proposed to leverage the important information from social relationships to enhance social relation atmosphere recognition. In this way, the recognition of subjective attributes is increased by importing useful objective information. The fused features are more representative than features directly extracted from end-to-end neural networks.

Chapter 5 introduces a spatial-temporal concept-based explanation method for explaining neural networks. Currently, video-based explanation methods mainly focus on pixel-level interpretation. None of them are able to produce a high-level explanation. An STCE (Spatial-Temporal Concept-based Explanation) framework is proposed for interpreting 3D Convolutional Neural Networks (CNNs) and explaining them by introducing human-understandable concepts. The concepts are grouped by supervoxels extracted from videos. The framework evaluates the importance score for each concept. The high score represents the network that pays more attention. The proposed framework is utilized in Chapter 4 as a plug-in module to help recognize the subjective attributes.

Chapter 6 gives the summary and prospect of this thesis.