Research class - workshop, 21.05.2021.

U petak 21.05.2021 održat će se workshop u sklopu Research class predavanja povezan s istraživanjima u sklopu HRZZ projekta **RAASS: Automatsko raspoznavanje akcija i aktivnosti u multimedijalnom sadržaju iz domene sporta**, voditeljice izv. prof. Marine Ivašić-Kos.

Suradnici na projektu i doktorski studenti predstavit će trenutne istraživačke aktivnosti i rezultate. Workshop će biti organizirati u hibridnom okruženju, u učionici S32 i putem portala Teams/ Doktorski studij.

Predviđen raspored:

Petak, 21.05.2021

10:00 - 10:30 **Slobodan Ribarić** (gost predavač):
Pristup utemeljen na znanju za analizu mnoštva ljudi u nadzornim sustavima

10:30 - 11:00 **Kristina Host**: Action recognition in handball scenes
11:00 - 11:30 **Romeo Šajina**: Pose estimation, tracking and comparison
11:30 - 12:00 **Matija Burić**: Knowledge based action recognition

(Coffee break)

13:00 - 13:30 **Miran Pobar**: Spatiotemporal action localization in handball video
13:30 - 14:00 **Goran Paulin**: Person localization and distance determination using raycast
14:00 - 14:30 **Saša Sambolek**: Transfer learning for training person detector in drone imagery
14:30 - 15:00 **Ingrid Hrga**: Visual Attention for Image Caption Generation

Croatian Science Foundation fully supports this research under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS).

Sažetak predavanja 21.05.2021, prostorija S32, virtualno Teams/Doktorski:

1. Pristup utemeljen na znanju za analizu mnoštva ljudi u nadzornim sustavima

Speaker: red. prof. **Slobodan Ribarić**, Fakultet elektrotehnike i računarstva Sveučilišta u Zagrebu

Abstract: Jedno od trenutno najaktivnijih istraživačkih područja računalnog vida je analiza scena s mnoštvom ljudi (engl. crowd). Mnoštvo se može definirati kao grupa ljudi ili velika skupina pojedinaca okupljenih u istom fizičkom okruženju. Modeliranje mnoštva, analiza scena s mnoštvom i raspoznavanje ponašanja su najzahtjevnije teme računalnog vida i umjetne inteligencije. Znanstveni cilj predloženog istraživanja je razvoj eksperimentalnog inteligentnog sustava utemeljenog na znanju za analizu mnoštva i raspoznavanje ponašanja na temelju vizualnih informacija dobivenih video nadzornim sustavom. Predložen je novi pristup utemeljen na znanju za modeliranje scena s mnoštvom, koji predstavlja spoj metoda utemeljenih na agentima i entitetima, i zdravorazumskog ljudskog znanja. Ideja je pretvoriti ljudsko (ili ekspertno) znanje o ponašanju mnoštva, utemeljeno na vizualnim informacijama, i dodati ga drugim modelima temeljenim na informacijama dobivenim metodama računalnog vida, u bazu znanja koju podržava stroj za zaključivanje. Glavni ciljevi istraživanja su: i) Kritički pregled i analiza ranijih pristupa na području modeliranja i analize mnoštva u sustavima video nadzora; ii) Novi model mnoštva koji se temelji na fuziji agenata i entiteta; iii) Razvijati metode izlučivanja značajki iz videa prilagođene za makroskopsku I mikroskopsku analizu scena s mnoštvom; iv) Pristup modeliranja mnoštva utemeljen na zdravorazumskom znanju; v) Nova arhitektura za hibridno predstavljanje znanja koja kombinira zdravorazumsko ljudsko znanje i koncepte dubokog učenja; vi) Razviti shemu hibridnog predstavljanja znanja te njezinu integraciju s dubokim neuronskim mrežama; vii) klasifikacija ponašanja mnoštva ljudi; viii) Dizajn dubokog modela učenja za analizu scena s mnoštvom; ix) Razvoj eksperimentalnog sustava i njegovo ispitivanje i evaluacija.

2. Action recognition in handball scenes

Speaker: Kristina Host, PhD student, Odjel za informatiku Sveučilišta u Rijeci

Abstract: Action recognition in sports, especially in handball, is a challenging task due to a lot of players being on the sports field performing different actions simultaneously. Training or match recordings and analysis can help an athlete, or his coach gain a better overview of statistics related to player activity, but more importantly, action recognition and analysis of action performance can indicate key elements of technique that need to be improved. In this paper the focus is on recognition of 11 actions that might occur during a handball match or practice. We compare the performance of a baseline CNN-model that classifies each frame into an action class with LSTM and MLP based models built on top of the baseline model, that additionally use the temporal information in the input video. The models were trained and tested with different lengths of input sequences ranging from 20 to 80, since the action duration varies roughly in the same range. Also, different strategies for reduction of the number of frames were tested. We found that increasing the number of frames in the input sequence improved the results for the MLP based model, while it didn't affect the performance of the LSTM model in the same way.

Croatian Science Foundation fully supports this research under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS).

3. Pose estimation, tracking and comparison

Speaker: Romeo Šajina, PhD student, Fakultet Informatike, Sveučilište J. Dobrile Pula

Abstract: Deep learning has become the number one research field with a lot of effort invested in computer vision. By providing possibilities as object detection, object tracking, and scene annotation, computer vision has found a lot of applications in the real world. In the field of sports, computer vision can be used to detect players, track players, detect and track the ball, detect player actions, detect objective score change, player pose estimation, etc. In this paper, we will describe player pose estimation, tracking, and comparison. This is particularly interesting as we can collect poses of a player executing an action (e.g., jump shot) and use it as a template for other players. By comparing other player's poses to the template poses we can provide them with the information of the needed corrections to their action execution.
This information can help players to improve their overall action execution sequence where they can evaluate their pose in each video frame. The closes that the action sequence is to the template sequence, the better the score will they achieve. This application can be especially useful to beginners in the sport, as later on in the career a top player can develop their style of executing certain action sequences, thus trying to correct them might compromise their performance. In this paper, we will discuss player pose tracking while executing an action sequence with pose comparison and evaluation techniques.

4. Knowledge based action recognition

Speaker: Matija Burić, PhD student, HEP d.d., Sektor za informacijsko-komunikacijske tehnologije, PS Rijeka

Abstract: The main focus of recent studies is based on approaches where machine learning is applied on datasets of images or frames in a video clip. This is expected considering how successful deep learning methods are compared to more conventional methods. By using neural networks, one is capable of better dealing with computer vision noise and clutter which all real-world datasets have in common. We are at the point where machines are more competent to recognize certain features of significance than humans are. For a human to detect an individual object in the image some higher level of semantic is needed for every new layer of complexity. For example, a human can easily recognize and count a small number of sports balls covering just a few cubical meters of a playfield but has difficulties if he needs to count all which cover the whole playing field. He will use mathematical instruments to make an assumption of how much individual object covers the floor area rather than count every single one like machines are capable of. That said we can conclude that machines are superior to humans but what if, for an acceptable detection, we need a set of rules which needs to be followed. What if we combine human knowledge and machine capabilities? Could we, by simple based rules, improve detection further or even filter good from bad detections at the beginning of detection while significantly decreasing resources machine learning needs. This is something to think about.

Croatian Science Foundation fully supports this research under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS).

5. Spatiotemporal action localization in handball video

Speaker: doc. dr. sc. **Miran Pobar**, Odjel za informatiku Sveučilišta u Rijeci

Abstract: Spatiotemporal action localization is the problem of identifying the spatiotemporal location where the action occurs, as well as assigning the appropriate action class to the identified segment. As the task pertains to temporally untrimmed videos, it poses additional challenges compared with action recognition, such as detecting whether a segment includes an action at all, detecting multiple actions occurring simultaneously and dealing with actions of different durations. However, the task is essential for analysis of realistic sports videos of whole matches or training sessions.
Here we present an approach for spatiotemporal action localization in handball videos that includes player detection, tracking and action recognition. For the action recognition part, we use the 3d CNN-based models trained on a custom dataset of isolated handball actions in 10 action classes. The proposed approach was tested on both the action recognition task in isolation on trimmed videos, and on a dataset of untrimmed handball training videos for the spatiotemporal action localization task.

6. Person localization and distance determination using raycast

Speaker: Goran Paulin, PhD student, Kreativni odjel d.o.o., Rijeka

Abstract: By using drones in search and rescue (SAR) missions, missing person detection is possible during or after the flight by analyzing the recorded material. However, person localization is equally important so that rescuers can approach the person in the shortest possible time. We propose a raycast method to determine both a person's location and the distance from the drone, using a sequence of monocular drone images. The proposed method has been tested in silico, using a custom-made procedural simulator, calibrated for windless and windy conditions. We concluded that multiple raycasting solves unreliable telemetry problems and that there is an optimal number of required iterations, depending on the telemetry noise of a specific drone.

7. Transfer learning for training person detection in drone imagery

Speaker: Saša Sambolek,PhD student, Srednja škola Tina Ujevića, Kutina

Abstract: Deep neural networks achieve excellent results on various computer vision tasks, but learning models require large amounts of tagged images and often unavailable data. An alternative solution of using a large amount of data to achieve better results and greater generalization of the model is to use previously learned models and adapt them to the task at hand, known as transfer learning.
The aim of this paper is to improve the results of detecting people in search and rescue scenes using YOLOv4 detectors. Since the original SARD data set for training human detectors in search and rescue scenes are modest, different transfer learning approaches are analyzed. Additionally, the VisDrone data set containing drone images in urban areas is used to increase training data in order to improve person detection results.

Croatian Science Foundation fully supports this research under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS).

7. **Visual Attention for Image Caption Generation**

Speaker: Ingrid Hrga, PhD student**,** Fakultet Informatike, Sveučilište J. Dobrile Pula

Abstract: In the world we live in, we are constantly exposed to large amounts of various stimuli. But despite that, we manage to solve everyday tasks in a focused way and to behave coherently. This is because only a fraction of the multitude of sensory information gets selected for detailed processing, while the rest remains neglected. Attention is a brain mechanism responsible for the selective processing only of the most relevant subsets of the sensory data. And in efforts to provide machines with human-like abilities, the attention mechanism has proven to be one of the key components.

The presentation consists of two parts. The first part discusses the biological basis of perception and attention. We explain some general concepts needed to better understand attentional models used in computer vision. In the second part, we draw a parallel between human and machine attention. Although the attention mechanisms are employed successfully in a variety of research areas, we focus on the specific task of image captioning that connects computer vision and natural language processing. We provide an overview of attention-based image captioning models, from the early ones, based on soft-attention to today's state-of-the-art transformer-based systems. We show how different approaches have contributed to the progress on a task that has long been considered extremely difficult for machines and is now close to be solved.

Croatian Science Foundation fully supports this research under the project IP-2016-06-8345 "Automatic recognition of actions and activities in multimedia content from the sports domain" (RAASS).