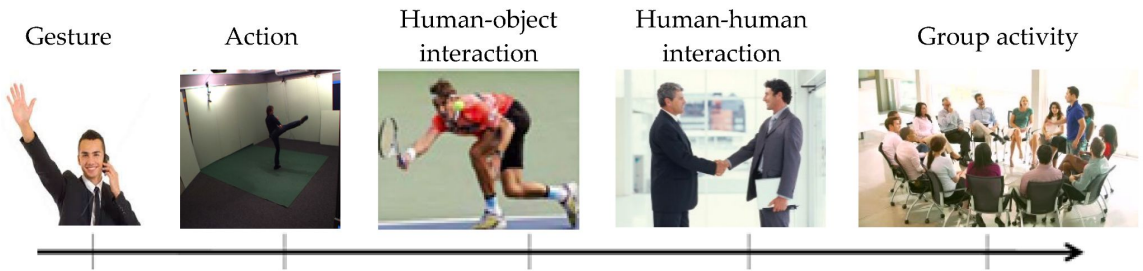
Rozpoznawanie dyscypliny sportu na podstawie materiału wideo

Rafał Lewanczyk
Opiekun: dr. inż. Artur Wilkowski

Wprowadzenie

Rozpoznawanie akcji można podzielić na kilka rodzai:

- klasyfikacja akcji (analiza fragmentu wideo zawierającego pojedynczą akcję)
  - klasyfikacja akcji wykonywanej przez 1 osobę
  - klasyfikacja akcji grupowej (można wydzielić pojedyncze osoby, wchodzące ze sobą w interakcje) np. bójka, sporty drużynowe
  - klasyfikacja akcji tłumu (nie da się wydzielić pojedynczych osób) np. demonstracja
  - wykrywanie akcji (znalezienie segmentu wideo zawierającego akcje, następnie sklasyfikowanie go)

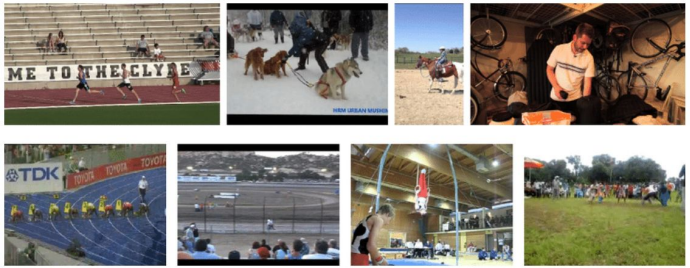


Wyzwania

- W grupowej klasyfikacji aktywności, akcje pojedynczych osób są niejednoznaczne np. stanie w kolejce, jest czymś innym niż staniem w grupie i rozmawianiem
- Główną informację może nieść niewielka ilość osób w grupie będąca otoczona innymi osobami niosącymi mniejszą ilość informacji lub zakłócającymi całkowity obraz. Należy rozważyć cały kontekst sytuacji. np. w meczu siatkówki atakująca oraz blok ruszają się bardzo dynamicznie, a reszta zawodników stoi w gotowości
- Modelowanie akcji wykonywanej przez człowieka
- Reprezentacja cech modelu (cechy reprezentujące wygląd i poza człowieka nie wystarczają, zamodelować należy również ich zmianę w czasie. Dodany zostaje kolejny wymiar reprezentujący czas)

Zbiory danych

- Broadcats Field Hockey Dataset - 58 nagrań z akcjami pojedynczych osób (np. podanie, kiwanie, strzał) oraz klasyfikacją całej akcji (obrona, atak itd.)
- Volleyball Dataset - 4830 nagrań z 55 meczy. Podzielony w taki sam sposób
- C-Sports - 11 różnych dyscyplin sportu min. koszykówkę, zbijaka, piłke nożną. W każdym nagraniu zawarty również opis akcji
- NBA Dataset - 9172 nagrań z 181 gier NBA. Zawiera opisy aktualnej akcji składające się z 9 aktywności.
- Sports-1M - Ponad 1 milion nagrań z serwisu YouTube podzielonych na 487 kategorii wg. etykiet nadanych przez serwis YouTube
- **Sports videos in wild (SVW) - Zbiór składa się z 4200 nagrań podzielonych na 30 dyscyplin sportowych. Nagrania pokazują sporty uprawiane przez profesjonalistów jak i amatorów.**



https://paperswithcode.com/dataset/sports-1m

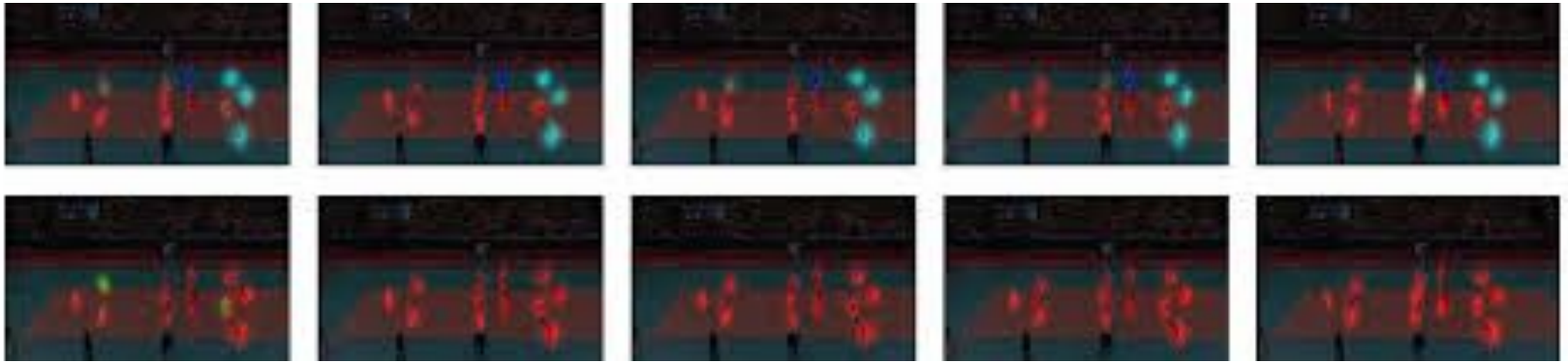
Cechy zawarte w danych

- dane RGB
  - cechy:
    - czasoprzestrzenne objętościowe cechy (spatiotemporal volume-based features)
    - czasoprzestrzenne cechy punktów zainteresowań
    - cechy śledzenia stawów
  - ograniczenia:
    - ruch kamery
    - ograniczenia detekcji człowieka
  - dane o głębi RGBD (wymaga specjalnych kamer)
    - cechy takie same jak RGB
    - zastosowanie głębi eliminuje problemy związane ze zmianami w otoczeniu
- Deep learning
  - model sam wyucza się najlepszych cech (cechy te w dużej mierze są dopasowane do realizowanego zadania, są trudne lub niemożliwe do zinterpretowania przez człowieka)

Metody rozpoznawania akcji przy ręcznym modelowaniu cech

Metody z góry do dołu (Top-Down Approach)

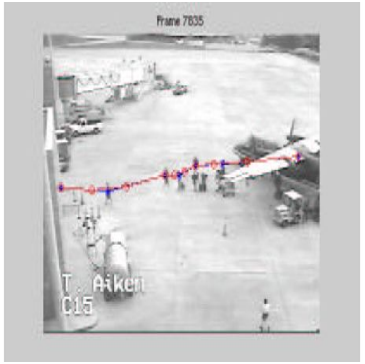
Skupiona na analizie globalnych wzorców ruchu całej grupy oraz śledzenie trajektorii, oraz interakcji w całej grupie. Akcje pojedynczego uczestnika w grupie są mniej ważne.



Metody oparte na trajektoriach

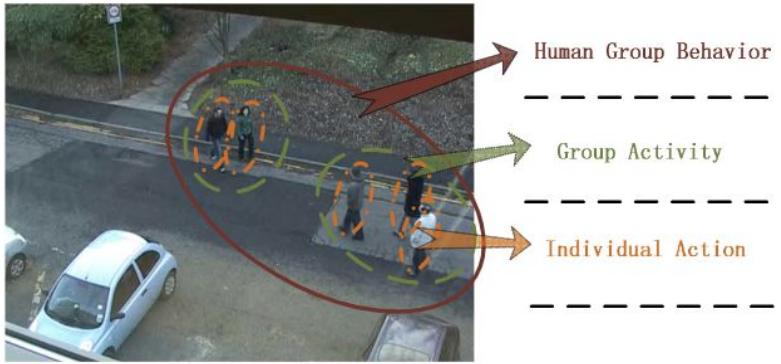
Analiza grupowych aktywności pod względem interakcji pomiędzy pojedynczymi trajektoriami.

Vaswani[5] modeluje zachowania grupowe poprzez reprezentowanie ruszających się obiektów jako punkty w dwuwymiarowej płaszczyźnie. W ten sposób grupowe zachowanie przedstawione jest jako zmieniający się w czasie wielokąt. Akcja zostaje przedstawiona jako średnia ze wszystkich kształtów w przestrzeni tangensowej.



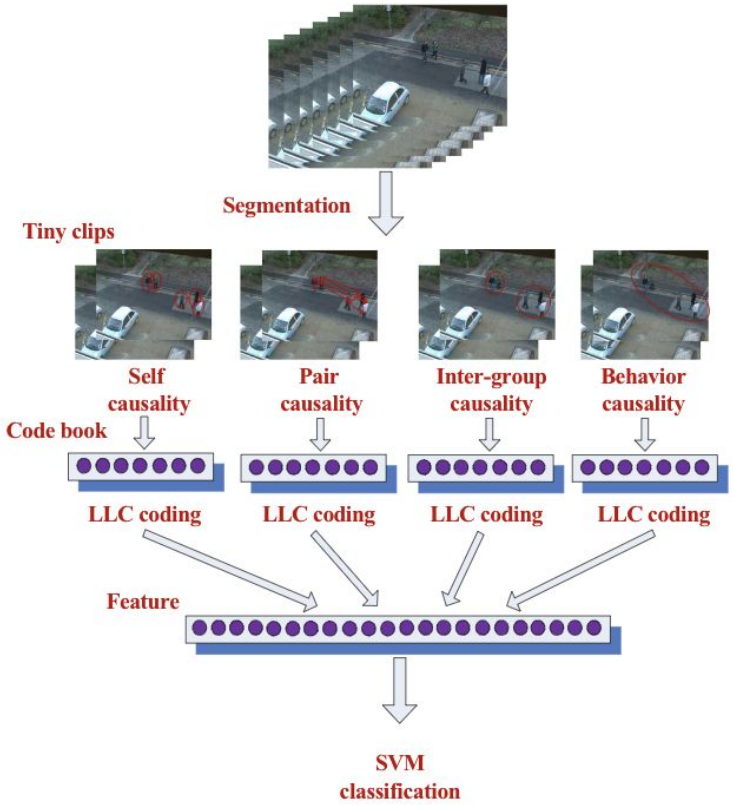
Metody oparte o podgrupy

W celu uproszczenia sytuacji zaproponowane zostały metody polegające na podział grupy wykonującej aktywność na podgrupy oraz szukania zależności pomiędzy podgrupami. Podział na podgrupy ułatwia zadanie klasyfikacji, jednak dokonanie poprawnego podziału jest bardzo trudne.



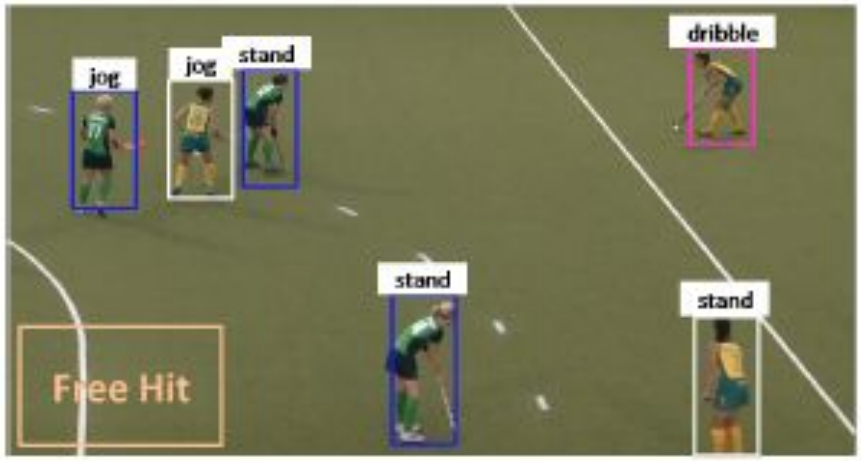
Metody oparte o podgrupy

Zhang et al. [2] zaproponował podział na podgrupy na podstawie wielo grupowej przyczynowości.
Wyznaczony został podział na udziałowość indywidualną, w parze, między-grupową, ogólną.
Każdy z podziałów został zakodowany przy pomocy kodowania
LCC, budując zbiór cech modelu SVM.



Metody z dołu do góry (Down-Top Approach)

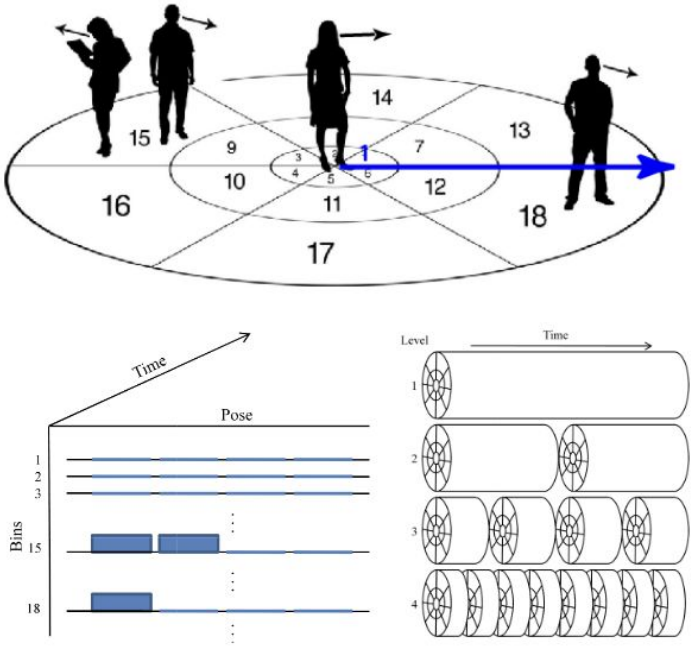
Skupiona na rozpoznawaniu pojedynczych osób następnie analizowanie ich struktury hierarchalnej na poziomie pojedynczej osoby oraz poziomie grupy.



Metody oparte o deskryptory

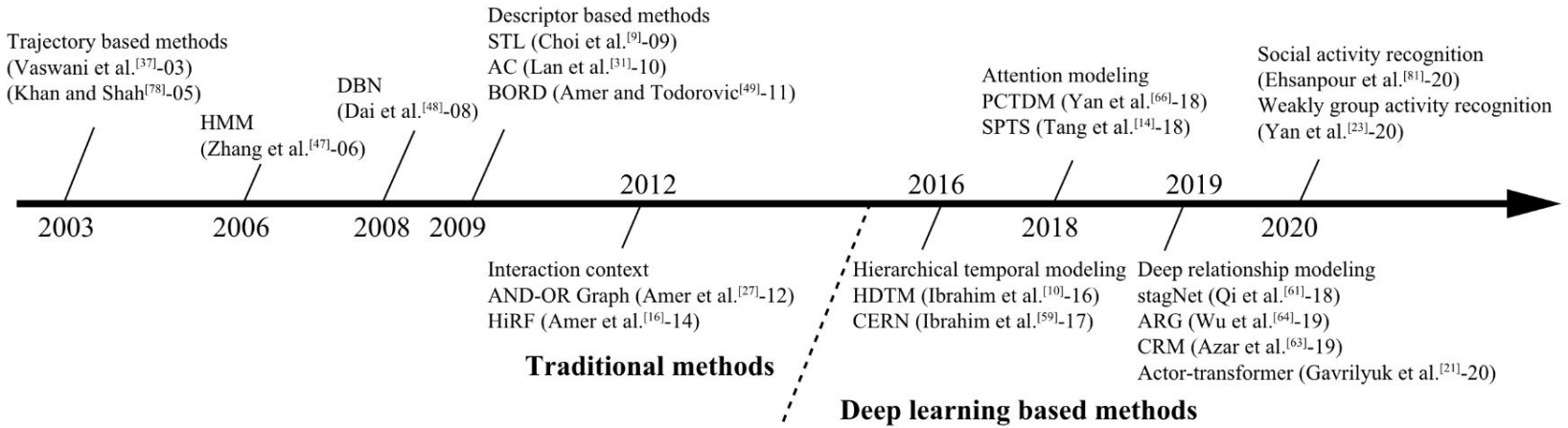
Metody te łączą ze sobą konteksty sceny poprzez deskryptory wyznaczające kluczowe elementy scen.

Choi et al.[10] przedstawia lokalnie czasowo-przestrzenny deskryptor (STL), który określa czasowo-przestrzenną dystrybucję pozycji, pozy, oraz informacji o ruchu pojedynczych osób. Osoby oraz ich sylwetki są wykrywane poprzez zastosowanie histogramów zorientowanych gradientów (HOG) oraz modelu SVM.



Metody Deeplearningu (automatyczna nauka cech)

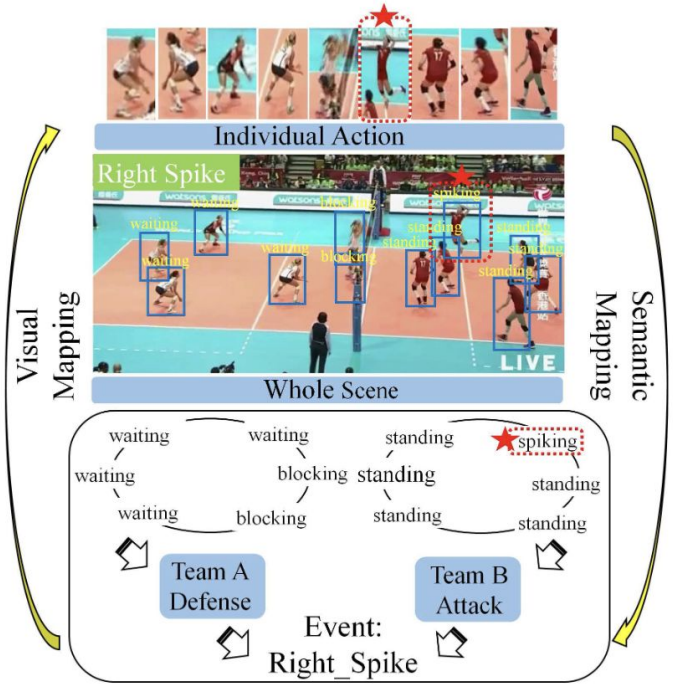
W ostatnich latach splotowe sieci neuronowe (CNN) osiągnęły bardzo dobre wyniki w zadaniach percepcji maszyn, takich jak rozpoznawanie obrazu, rozpoznawanie elementów na obrazie lub klasyfikacja wideo. Po zastosowaniu sieci CNN w zadaniach klasyfikacji aktywności grupowej osiągnęły one znacznie lepsze wyniki od metod używających ręcznie wyznaczone cechy modelu.



Głębokie modelowanie relacji grupowych

Znajdowanie związków pomiędzy akcjami uczestników aktywności jest kluczowe dla dobrej klasyfikacji. Znajdowanie tego typu relacji jest skomplikowane ze względu na brak dostępnych danych, zbiory treningowe zawierają albo informacje dotyczące aktywności pojedynczych osób w grupie lub całej grupy. Dużo badań poszukuje sposobu na odnalezienie metody znajdowania relacji pomiędzy uczestnikami aktywności.

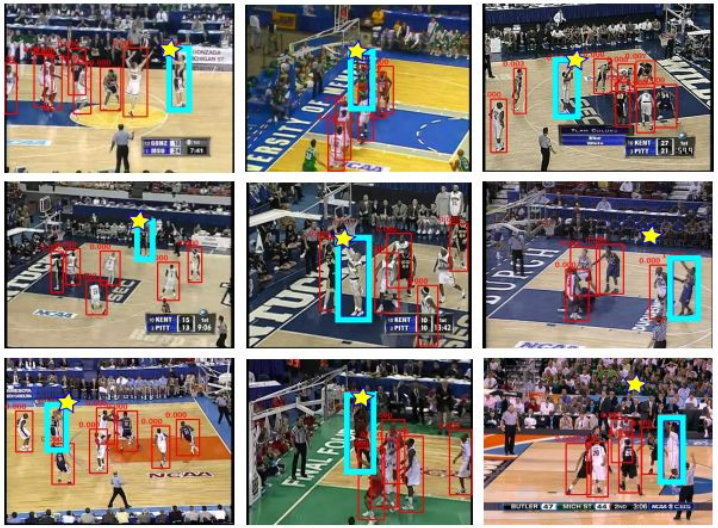
Qi et al. [6] zaproponował stagNet - rekurencyjną sieć neuronową wyposażoną w atencję. Rozwiązanie buduje graf semantyczny, używając etykiet słownych oraz danych wizualnych. Relacja przestrzenna pomiędzy pojedynczymi uczestnikami akcji jest wnioskowana w grafie semantycznym na podstawie mechanizmu przekazywania wiadomości. Ponad to model śledzi kluczową osobę w akcji.



Modele z uwagą

W większości aktywności najwięcej informacji niesie jedna lub kilka osób wykonywujących akcję, pozostałe osoby mogą powodować zakłócenie modelu. Ze względu na brak informacji w zbiorach danych określających kluczową osobę, problem ten jest dużym wyzwaniem.

Ramanathan et al. [7] zastosował model z czasową uwagą na zbiorze danych składającym się z meczów z koszykówki, znajdując w scenie kluczowych zawodników i poprawiając odczyt aktualnego zagrania.



Hierarchiczne modelowanie czasowe

W pracy decyduje się na modyfikację podejścia zaprezentowanego w pracy A Hierarchical Deep Temporal Model for Group Activity Recognition

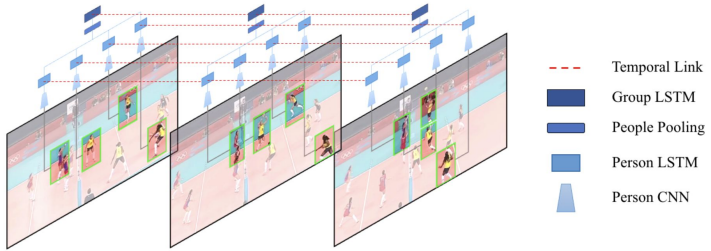
Podejście to charakteryzuje się zastosowaniem dwóch warstw LSTM:

Pierwsza warstwa składa się z kilku LSTM odpowiada za śledzenie akcji pojedynczej osoby. Na jej wejściu znajdują się cechy wydobyte przez sieć CNN z bounding box'a badanej osoby w danym momencie.

Druga warstwa przyjmuje na wejściu połączone wyjścia z pierwszej warstwy. Swoje wyjście kieruje do warstwy klasyfikującej soft-max, wykonującej predykcję aktywności grupowej.

Model ten był trenowany w 2 krokach. Jako pierwsza wytrenowana została sieć LSTM odpowiedzialna za akcje pojedynczej osoby. Druga warstwa została wytrenowana przy sprawnie działającej pierwszej warstwie.

Model osiągnął skuteczność 81.5% na zbiorze z ogólnymi aktywnościami Collective Activity Dataset.



Realizowane podejście

W pracy magisterskiej, również planuje zbudować 2-warstwowy model, składający się z sieci LSTM. Warstwy pełnią takie same funkcje co w podejściu wzorcowym, jednak będą one przybierały inne parametry na swoim wejściu:

Pierwsza warstwa LSTM na wejściu przyjmuje punkty kluczowe wykrytej osoby, odpowiadające w jakiej pozycji się znajduje.

Druga warstwa LSTM przyjmuje na wejściu połączone wyjścia z pierwszej warstwy. Przy eksperymentach nad tą warstwą, możliwe, że jej wejście zostanie rozszerzone o dodatkowe parametry np. cechy charakterystyczne obrazu danej klatki.

Przegląd bibliotek ekstrakcji punktów kluczowych

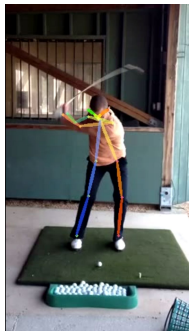
- OpenPose (https://github.com/CMU-Perceptual-Computing-Lab/openpose)
- **AlphaPose (https://github.com/MVIG-SJTU/AlphaPose)**
  - **wykrywanie poz**
  - **śledzenie poz**

Wstępny eksperyment cz. 1

Zrealizowany został pierwszy wstępny eksperyment treningu pierwszej warstwy LSTM. Eksperyment składał się z następujących kroków:

1. Ze zbioru SVW wybrane zostały 3 klasy sportów indywidualnych (łucznictwo, tenis, golf) oraz po 40 nagrań z każdej z tych klas.

2. Z każdego z nagrań wyeksportowane zostały pozy sportowca

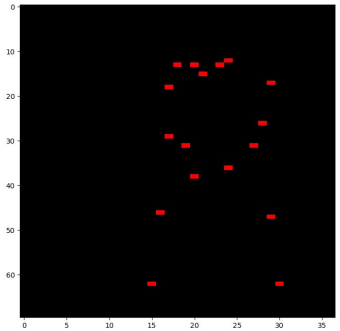


Wstępny eksperyment cz. 2

Opracowana została metoda wstępnego przetworzenia danych zawierających pozę uczestnika na formę pozwalającą użycie ich na wejściu modelu LSTM.

Przetworzenie składa się z następujących kroków:

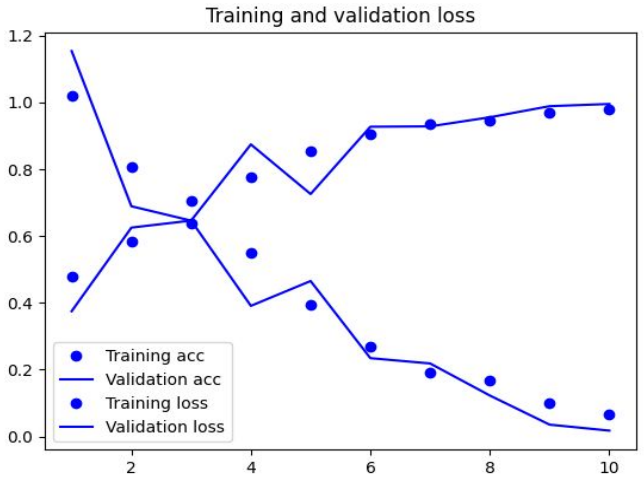
a. Przesunięcie pozy do punktu (0, 0), w celu usunięcia poruszania się obiektu lub kamery
b. Ograniczenie otoczenia pozy do jej największego bounding box'a
c. Normalizacja punktów kluczowych pozy w ramach wybranego bounding box'a



Wstępny eksperyment cz. 3

Opracowany został podstawowy model LSTM klasyfikujący indywidualną aktywność sportową. Model osiągnął skuteczność klasyfikacji 94%.

Kod eksperymentu został umieszczony w repozytorium



Dalsze prace

Na dalsze prace składają się:

1. Opracowanie filtra usuwającego sylwetki nie wnoszących informacji np. wykrytą widownię
2. Rozszerzenie modelu z pierwszego eksperymentu na nagrania, w których bierze udział więcej niż jedna osoba
3. Opracowanie drugiej warstwy LSTM klasyfikującą aktywność grupową.

Bibliografia

[1] Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. In *Sensors (Switzerland)* (Vol. 19, Issue 5). MDPI AG. https://doi.org/10.3390/s19051005

[2] Zhang, C., Yang, X., Lin, W., & Zhu, J. (2012). Recognizing human group behaviors with multi-group causalities. *Proceedings of the 2012 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, WI-IAT 2012*, 44–48. https://doi.org/10.1109/WI-IAT.2012.162

[3] Yan, S., Xiong, Y., & Lin, D. (2018). *Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition*. http://arxiv.org/abs/1801.07455

[4] Wu, L. F., Wang, Q., Jian, M., Qiao, Y., & Zhao, B. X. (2021). A Comprehensive Review of Group Activity Recognition in Videos. In *International Journal of Automation and Computing* (Vol. 18, Issue 3, pp. 334–350). Chinese Academy of Sciences. https://doi.org/10.1007/s11633-020-1258-8

[5] Vaswani, N., Chowdhury, R., & Chellappa, R. (n.d.). *Activity Recognition Using the Dynamics of the Configuration of Interacting Objects*.

[6] Qi, M., Wang, Y., Qin, J., Li, A., Luo, J., & van Gool, L. (2020). StagNet: An Attentive Semantic RNN for Group Activity and Individual Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(2), 549–565. https://doi.org/10.1109/TCSVT.2019.2894161

[7] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, F. F. Li. Detecting events and key actors in multi-person videos. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, USA, pp.3043−3053, 2016. DOI: 10.1109/CVPR. 2016.332.

[8] Lan, T., Wang, Y., Yang, W., Robinovitch, S. N., & Mori, G. (2012). Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *34*(8), 1549–1562. https://doi.org/10.1109/TPAMI.2011.228

[9] Ibrahim, M., Muralidharan, S., Deng, Z., Vahdat, A., & Mori, G. (2015). *A Hierarchical Deep Temporal Model for Group Activity Recognition*. http://arxiv.org/abs/1511.06040

[10] Choi, W., Shahid, K., & Savarese, S. (2009). What are they doing?: Collective activity classification using spatio-temporal relationship among people. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, 1282–1289. https://doi.org/10.1109/ICCVW.2009.5457461

[11] Cheng, Z., Qin, L., Huang, Q., Jiang, S., & Tian, Q. (2010). Group activity recognition by gaussian processes estimation. *Proceedings - International Conference on Pattern Recognition*, 3228–3231. https://doi.org/10.1109/ICPR.2010.789

[12] Azar, S. M., Atigh, M. G., & Nickabadi, A. (2018). *A Multi-Stream Convolutional Neural Network Framework for Group Activity Recognition*. http://arxiv.org/abs/1812.10328

[13] T. Lan, Y. Wang, G. Mori, S. N. Robinovitch. Retrieving actions in group contexts. In Proceedings of European Conference on Computer Vision, Springer, Heraklion, Greece, pp.181−194, 2010. DOI: 10.1007/978-3-642-35749-

[14] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, F. F. Li. Detecting events and key actors in multi-person videos. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, USA, pp.3043−3053, 2016. DOI: 10.1109/CVPR. 2016.332.