Dr inż. Artur Wilkowski
Instytut Automatyki i Informatyki Stosowanej
Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska
ORCID: 0000-0002-6814-7645
e-mail: artur.wilkowski@pw.edu.pl

METODY ROZPOZNAWANIA AKTYWNOŚCI W SEKWENCJACH WIDEO PRZY UŻYCIU NISKOPOZIOMOWYCH CECH OBRAZU

Streszczenie

Zagadnienia związane z problemem rozpoznawania aktywności w materiałach wideo znajdują się w centrum zainteresowania wielu służb odpowiedzialnych za bezpieczeństwo publiczne. Odpowiednie systemy rozpoznające mogą zwiększać bezpieczeństwo obywateli poprzez wykrywanie zachowań niebezpiecznych, agresywnych i nielegalnych w monitoringu wizyjnym lub przez detekcję i filtrowanie niepożądanych i nielegalnych materiałów wideo dostępnych w Internecie, w tym materiałów pornograficznych lub materiałów CSAM. Dane ekstrahowane z wideo służące do rozpoznania mogą mieć charakter cech niskopoziomowych (np. kolor, ruch) lub wysokopoziomowych np. wykryte pozycje stawów sylwetek ludzkich. Artykuł podejmuje się zadania przedstawienia rozwiązań informatycznych umożliwiających klasyfikację materiałów wideo ze szczególnym uwzględnieniem wykorzystania cech niskopoziomowych. Przedstawione są metody klasyczne oraz najnowsze metody wykorzystujące uczenie głębokie.

Słowa kluczowe: rozpoznawanie aktywności ludzkiej, przetwarzanie wideo, sieci neuronowe.

WSTĘP

Ze względu na intensywny rozwój technik akwizycji i sieciowej wymiany danych dostęp do materiałów wideo jest w dzisiejszych czasach niezwykle łatwy. Rozliczne materiały filmowe są dostępne w Internecie a nasze miasta są wypełnione kamerami wizyjnymi monitorującymi bezpieczeństwo. Specjaliści stale przeglądają dostępne materiały wideo pochodzące z monitoringu wizyjnego dbając o bezpieczeństwo publiczne, a materiały z Internetu oceniane są pod kątem legalności oraz odpowiedniości dla wybranych grup odbiorców. Ze względu na ograniczenia ludzkiej percepcji, człowiek może przetworzyć tylko stosunkowo niewielką liczbę materiałów wideo. Fakt, że liczba materiałów jest bardzo duża i tylko niewielki odsetek materiałów (szczególnie w przypadku monitoringu) stanowi podejrzane treści, dodatkowo zmniejsza efektywność takiej pracy.

Z tego powodu istnieje duża potrzeba rozwoju automatycznych systemów wspomagających człowieka w filtrowaniu materiałów wideo. Współcześnie, takie systemy opierają się na zaawansowanych metodach sztucznej inteligencji, uczenia

maszynowego i sieciach neuronowych. Przykładem takiego systemu wspomagającego jest np. system konstruowany w ramach projektu APAKT[1], wycelowany w wykrywanie treści CSAM (Child Sexual Abuse Material). Innym przykładem może być eksperymentalny system wykrywania zachowań agresywnych w materiałach wideo[2].

Użycie metod uczenia maszynowego w celu ewaluacji materiałów wideo wymaga ekstrakcji z poszczególnych klatek oraz ich sekwencji interesujących informacji, tzw. cech, które są istotne z punktu widzenia realizowanego zadania. Cechy takie mogą być ekstrahowane na podstawie wiedzy eksperckiej (hand-crafted features), ale mogą też być realizowane w sposób automatyczny (np. cechy sieci splotowych - CNN features). W niniejszej pracy zrealizowany zostanie przegląd metod rozpoznawania aktywności z dużym naciskiem na różne rodzaje cech obrazu wykorzystywane jako wejście klasyfikatorów. Badane będą przede wszystkim zastosowania cech o charakterze niskopoziomowym, w których nie analizuje się jawnie struktury i pozy sylwetki ludzkiej w ramach etapu pośredniego rozpoznawania.

W kolejnych sekcjach prezentowane są uniwersalne cechy sekwencji obrazów umożliwiające rozpoznawanie zróżnicowanych aktywności oraz odpowiednie systemy rozpoznawania z nich korzystające.

CECHY SEKWENCJI OBRAZÓW

W ramach artykułu przedstawiono analizę źródeł w zakresie wyboru cech przestrzenno-czasowych umożliwiających skuteczne rozpoznawanie aktywności osób w materiałach wideo. Metody reprezentacji cech wideo zostały (zgodnie z typowo przyjmowanymi kryteriami[3]) podzielone na:

- globalne cechy sylwetki
- deskryptory lokalne
- cechy uczone za pomocą sieci neuronowych

Prominentni reprezentanci poszczególnych kategorii są opisani w dalszej części tekstu.

GLOBALNE CECHY SYLWETKI

Cechy reprezentacji globalnej określane są dla obiektu jako całości. Praktyczne zawsze wymagana jest wstępna segmentacja obrazu, czyli oddzielenie obiektów pierwszego planu oraz tła, a następnie obliczane są cechy określające kształt obszaru w czasie, cechy konturu lub inne cechy opisujące obszar w poszczególnych klatkach (np. przepływ optyczny, HOG – Histogram of Oriented Gradients).

[1] Niewiadomska-Szynkiewicz E., Różycka M., Staciwa K., Nyczka K., „System wspomagający wykrywanie treści wizualnych i tekstowych zagrażających bezpieczeństwu dzieci w cyberprzestrzeni.” Cybersecurity and Law 2023, nr 2(10), s. 202-220, 2023.

[2] Adamiok F., Wilkowski A., „Comparison of deep learning approaches to violence detection in videos”, Progress in Polish Artificial Intelligence Research 5. Proceedings of the 5th Polish Conference on Artifcial Intelligence (PP-RAI’2024) 18–20.04.2024, Warsaw, Poland, ed. Mańdziuk Jacek, Żychowski Adam, Małkiński Mikołaj (red.), Politechnika Warszawska, s. 249-256.

[3] Zhang S., Wei Z., Nie J., Huang L., Wang S. & Li Z. „A Review on Human Activity Recognition Using Vision-Based Method. Journal of Healthcare Engineering”. s. 1-31, 2017.

Jednym z prostszych sposobów reprezentacji danych zadaniu rozpoznawania aktywności są wzorce binarne[4]. W cytowanej pracy badane są metody rozpoznawania osób na podstawie sposobu ich poruszania się. Kolejne klatki sekwencji przetwarzane są za pomocą metod usuwania tła. Realizowane jest to przez porównanie ze sobą kolejnych klatek obrazu i obserwacje zmian poszczególnych pikseli. W ten sposób można wyodrębnić sylwetki poruszających się osób. Sylwetki te są następnie reprezentowane przez wektory binarne. W celu porównania tych wektorów z wektorami wzorcowymi wykorzystywane są proste metryki porównujące obszary np. odległość Euklidesa, znormalizowana korelacja, czy suma różnic bezwzględnych (SAD). Cechy są obliczane oddzielnie dla poszczególnych klatek, za interpretację ruchu odpowiadają algorytmy wyższego poziomu (HMM-Ukryte Modele Markowa).

Zamiast wzorców binarnych możliwe jest również wykorzystanie informacji konturowej. Kontur osoby jest opisywany np. za pomocą sekwencji punktów kontrolnych (tzw. reprezentacja Kendalla)[5]. W cytowanym artykule badane są zarówno cechy opisujące kształt sylwetki oraz cechy odwołujące się wyłącznie do kinematyki ruchu. W pierwszym przypadku w celu analizy sekwencji wykorzystywane są modele dopasowania sekwencji czasowych oparte o marszczenie czasu (DTW) oraz Ukryte Modele Markowa(HMM), w drugim badane są parametry ruchu o powtarzalnym charakterze (modele autoregresywne AR i ARMA). Bezpośrednie dopasowanie współrzędnych konturów osób w poszczególnych klatkach jest trudne ze względu na potencjalne zróżnicowane skale, położenie i rotacje konturów. Z tego powodu zastosowane są zaawansowane metody porównywania konturów odporne na takie zjawiska (opierające się o tzw. odległości Prokrusta).

Alternatywną metodą opisu sylwetki może być deskryptor PCA-HOG[6]. Deskryptor oparty jest o klasyczny deskryptor HOG obszaru. W deskryptorze tym analizowany jest obszar obrazu podzielony na komórki. W ramach każdej z komórek badane są zmiany jasności obrazu (gradienty), ich siła i kierunki. Taka informacja jest następnie agregowana za pomocą histogramu oraz normalizowana. Ze względu na duży rozmiar opisu cech stosowana jest redukcja wymiaru PCA, której celem jest zachowanie jak największej ilości informacji z oryginalnej przestrzeni cech przy jednoczesnej redukcji rozmiaru wektora. Proponowany deskryptor jest użyty zarówno do śledzenia obszaru, jak również do rozpoznania aktywności z wykorzystaniem algorytmów wyższego poziomu (Ukryte Modele Markowa - HMM).

Interesującym sposobem opisu aktywności jest przedstawienie jej w postaci śladu ruchu sylwetki w czasie[7]. W podejściu tym poszczególne klatki sekwencji są obrazami binarnymi. Ruch osoby jest agregowany w czasie i reprezentowany przez

[4] Sundaresan A., RoyChowdhury A., Chellappa R., „A hidden Markov model based framework for recognition of humans from gait sequences,", Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429), Barcelona, Spain, s. II-93, 2003.

[5] Veeraraghavan A., Chowdhury A. R., Chellappa R., „Role of shape and kinematics in human movement analysis", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, pp. I–730, Washington, DC, USA, 2004

[6] Lu W.-L., Little J. J., „Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor," The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), Quebec, Canada, s. 6-6, 2006.

[7] Bobick A. F., Davis J. W., „The recognition of human movement using temporal templates," w IEEE Transactions on Pattern Analysis and Machine Intelligence, tom: 23(3), s. 257-267.

pojedynczą ramkę zawierającą dane o ruchu z wielu klatek. Dwa podstawowe rodzaje tego podejścia do agregacji to Motion Energy Image (MEI) oraz Moment History Image (MHI). W obu przypadkach dla każdej pary sąsiednich klatek generowany jest obraz różnicowy, który jest następnie progowany w celu określenia obszarów ruchu. W ujęciu pierwszym (MEI) obszary ruchu są w prosty sposób agregowane przez sumowanie w określonym oknie czasowym, w przypadku MHI za pomocą skali szarości reprezentowany jest nie tylko fakt ruchu, ale czas jaki minął od jego końca. W opisanym podejściu obrazy MEI i MHI są opisywane za pomocą cech geometrycznych (momenty Hu) i porównywane za pomocą metryki (w proponowanym rozwiązaniu - odległości Mahalanobisa).

Innym rodzajem cech często wykorzystywanym do kodowania informacji o ruchu jest przepływ optyczny. W celu obliczenia przepływu optycznego porównuje się ze sobą kolejne klatki obrazu i stara się ustalić ruch któremu podlegają punkty obrazu. W ten sposób powstaje obraz o rozmiarach oryginalnego obrazu, którego piksele opisywane są przez wektory ruchu poszczególnych punktów w obrazie. Efros, Berg, Mori i Malik[8] proponują rozwiązanie, w którym postać, której aktywność jest rozpoznawana, jest śledzona za pomocą metod korelacyjnych, natomiast wektory cech dla postaci w kolejnych klatkach są ekstrahowane na podstawie danych przepływu optycznego obliczonego metodą Lucasa-Kanade. Obliczony przepływ jest dzielony na kategorie (lewo, prawo, góra, dół) oraz rozmywany przy użyciu filtru Gaussa. Deskryptorem ruchu są te 4 kanały przepływu optycznego obliczone dla kilku punktów czasowych w sąsiedztwie przetwarzanej ramki obrazu. W proponowanej metodzie porównywane są ramki dwu sekwencji na zasadzie każdy-z-każdym, tworząc macierz podobieństw w której elementy o orientacji diagonalnej są następnie wzmacniane za pomocą odpowiedniego filtra (diagonalnego). Podobieństwa aktywności są ewaluowane na podstawie siły podobieństw odczytanych z macierzy podobieństw.

Interesującym sposobem reprezentacji ruchu osoby jest wykorzystanie wolumenu czasowo-przestrzennego[9], który utworzony jest przez obrys poruszającej się sylwetki. W cytowanym rozwiązaniu algorytm usuwania tła jest używany do detekcji pierwszego planu. Następnie w pewnym oknie czasowym uzyskane obrazy binarne są integrowane w wolumen. Tak utworzony wolumen jest normalizowany i opisywany przez zestaw 14 momentów geometrycznych stopnia co najwyżej 2. Dłuższe sekwencje są rozpoznawane przy użyciu algorytmów wyższego poziomu (za pomocą Ukrytych Modeli Markowa - HMM).

ANALIZA SEKWENCJI CZASOWYCH DLA CECH GLOBALNYCH

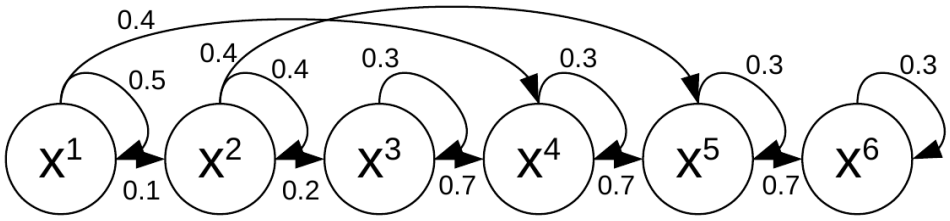
W zależności od przyjętego modelu cech, obliczone wektory bezpośrednio uwzględniają lub nie uwzględniają czasowy charakter danych. W pierwszym przypadku, w celu dokonania klasyfikacji wystarczające jest zwykle skorzystanie z uniwersalnego zestawu klasyfikatorów np. SVM, las losowy lub sieć neuronowa typu perceptron. Natomiast, jeżeli obliczone cechy są przypisane do poszczególnych

[8] Efros A.A., Berg A.C., Mori G. and Malik J., „Recognizing action at a distance," Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, tom. 2, s. 726-733, 2003.
[9] Achard C., & Qu X. & Mokhber A. & Milgram M. „A novel approach for recognition of human actions with semi-global features". Mach. Vis. Appl. tom. 19. s. 27-34, 2008.

klatek sekwencji (czyli tak naprawdę uzyskujemy sekwencję cech), muszą zostać zastosowane dodatkowe narzędzia modelowania temporalnego. Głównym problemem przy porównywaniu dwu sekwencji obrazów jest nierównomierny czas trwania poszczególnych etapów sekwencji pomiędzy przykładami (np. ktoś wykonuje dany etap czynności w różnym tempie niż inna osoba). Narzędzia modelowania muszą uwzględniać te zniekształcenia.

Jednym z narzędzi, które dobrze sprawdza się w porównywaniu sekwencji o nierównomiernym tempie wykonania akcji jest Dyskretne Marszczenie Czasu (Dynamic Time Warping – DTW)[10]. Narzędzie umożliwia dopasowanie do siebie poszczególnych klatek sekwencji w taki sposób, żeby całkowity koszt dopasowania był jak najmniejszy (tzn. dopasowane były klatki jak najbardziej podobne do siebie). Realizowane jest to poprzez dopuszczenie możliwości przypisywania kilku kolejnych klatek jednej sekwencji do pojedynczej klatki z drugiej sekwencji, w ten sposób kompensując różnice w tempie wykonania akcji. Tak dopasowane sekwencje mogą być łatwo porównane przy użyciu innych metod.

Alternatywnym i bardzo często wykorzystywanym narzędziem są Ukryte Modele Markowa (Hidden Markov Models – HMM)[11]. Są to modele probabilistyczne. Integralną częścią UMM, która odpowiada za kompensację nierównomierności w tempie wykonania akcji jest łańcuch Markowa (Rys. 1). Łańcuch Markowa składa się z stanów oraz prawdopodobieństw przechodzenia między nimi. Każdy ze stanów może reprezentować pewien etap wykonania akcji, prawdopodobieństwa przejścia między stanami określają oczekiwane tempo przechodzenia pomiędzy etapami, ale zapewniają też sporą elastyczność jeżeli chodzi o modelowanie różnic w tempie wykonania dla poszczególnych przykładów. W odróżnieniu od DTW model HMM nie porównuje bezpośrednio dwu sekwencji, natomiast porównuje nowe sekwencje z modelem uczonym na podstawie zbioru sekwencji.



Rys.1. Przykładowy łańcuch Markowa

DESKRYPTORY LOKALNE

W odróżnieniu od poprzednio opisywanych metod deskryptory lokalne nie próbują od razu opisać obszaru jako całości, lecz najczęściej zajmują się wykrywaniem i

[10] Myers C., Rabiner L., Rosenberg A., „Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," Acoustics, Speech, and Signal Processing, IEEE Transactions on, tom 28(6), s. 623–635, 1980.

[11] Rabiner L.R., „A tutorial on hidden Markov models and selected applications in speech recognition," Proceedings of the IEEE, tom 77(2), s. 257-286, 1989.

opisem punktów charakterystycznych obrazów i sekwencji obrazów. Częstym sposobem agregacji w czasie takich cech są metody typu Bag-of-Words.

Jedna z fundamentalnych metod w tej kategorii została zaproponowana przez Lapteva i Lindberga[12]. Badacze proponują metodę wykrywania interesujących punktów w sekwencji obrazów. W tym celu adaptowana jest metoda Harrisa, powszechnie wykorzystywana dla obrazów wyszukiwania punktów charakterystycznych statycznych obrazów. Metoda Harrisa koncentruje się na poszukiwaniu punktów, które są „interesujące" tzn. charakteryzuje je duża zmienność jasności w kierunkach obu osi (czyli głównie narożniki). W proponowanym rozwiązaniu poszukiwane są punkty o dużej zmienności nie tylko w przestrzeni obrazu, ale również w czasie (tzn. „narożniki" o dużej intensywności ruchu). Dodatkowo przeprowadzana jest przeszukiwanie w celu odnalezienia optymalnej skali narożnika. Dla tak określonych punktów (określanych STIP – Spatio-Temporal Interest Points) obliczane są zróżnicowane wektory cech składające się m.in. z:

- pochodnych funkcji jasności
- lokalnego przepływu optycznego
- agregatów przepływu optycznego oraz gradientów
- składowych głównych przepływu optycznego oraz gradientów

Porównanie sekwencji realizowane jest poprzez zachłanne dopasowanie punktów charakterystycznych na podstawie podobieństwa deskryptorów i metryce określonej na podstawie m najlepszych dopasowań.

Oikonomopoulos, Patras Pantic[13] w nieco inny sposób podchodzą do problemu ekstrakcji punktów charakterystycznych z wolumenu czaso-przestrzennego obrazów. Dla każdego punktu wolumenu obliczana jest (w różnych skalach) lokalna miara nieuporządkowania (entropia Shannona) sygnału w pewnym jego otoczeniu. Na jej podstawie wybierane są wyróżniające się punkty. Wykorzystywany następnie mechanizm łączenia (klasteryzacji) przekształca zbiory punktów w regiony. Jako cechy sekwencji wykorzystywane są tylko współrzędne x-y-t centroidów regionów, które tworzą zbiory. Takie zbiory mogą być porównywane za pomocą metryk odległościowych dla zbiorów (np. odległość Chamfera). Do prawidłowego porównania sekwencji wykorzystane są dodatkowo metody marszczenia czasu (DTW).

Oprócz nowych detektorów i deskryptorów dostosowanych do przetwarzania wolumenów czasowo przestrzennych proponuje się również 3-wymiarowe adaptacje istniejących detektorów/deskryptorów obrazów 2-wymiarowych takich jak SIFT oraz SURF. Dobrym przykładem jest deskryptor SIFT 3D[14]. W odróżnieniu do oryginalnej metody punkty charakterystyczne nie są wybierane, a losowane. Następnie dla każdego piksela sekwencji obrazów określana jest zmienność jasności w przestrzeni oraz w czasie, i takie wektory 3-wymiarowe są agregowane poprzez 3-

[12] Laptev I., Lindeberg, T. Local Descriptors for Spatio-temporal Recognition. ECCV, tom 3667. s. 91-103, 2004.

[13] Oikonomopoulos A., Patras I., Pantic M., „Spatiotemporal salient points for visual recognition of human actions," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), tom 36(3), s. 710-719, 2006.

[14] Scovanner P., Ali, S., Shah, M., „A 3-dimensional SIFT descriptor and its application to action recognition," Proceedings of the ACM International Multimedia Conference and Exhibition. S. 357-360, 2007.

wymiarowe histogramy (w układzie biegunowym) w sposób co do zasady zbliżony do deksryptora HOG (ale rozszerzonego do danych 3D). Obliczony histogram ma zatem charakter 2-wymiarowy w odróżnieniu od „zwykłego" SIFT-a, gdzie histogram ma jeden wymiar. Agregacja cech obdywa się za pomocą metody Bag-of-Words.

Podobne rozwiązanie prezentują Kläser i Marszalek[15]. Tutaj jednak wykorzystywane są punkty charakterystyczne STIP, a szybkie obliczenie gradientów uzyskane jest przez użyciu tzw. całkowego wideo i filtra prostokątnego.

Również deskryptor SURF doczekał się swojej wersji czasowo-przestrzennej[16]. W cytowanej pracy został położny nacisk na szybkość działania detektora oraz deskryptora. Reprezentacja przestrzenno-temporalno-skalowa jest tutaj uzyskana za pomocą zastosowania filtracji filtrem prostokątnym, który może być bardzo wydajnie obliczony za pomocą całkowego wideo. Następnie, podobnie jak w oryginalnym detektorze wyznacznik macierzy drugich pochodnych jest używany do określenia punktów zainteresowania, a sam deksryptor wykorzystuje transformaty falkowe Haara.

Pewnym zwieńczeniem koncepcji użycia deskryptorów opartych o punkty charakterystyczne STIP jest praca nie tylko rozszerzająca znany deskryptor HOG o dodatkowy wymiar, ale proponująca podobny deskryptor dla przepływu optycznego[17]. W cytowanym rozwiązaniu dla każdego z wykrytych w sekwencji punktów charakterystycznych STIP określane są dwa rodzaje deskryptorów. Pierwszy z nich - HOG (3D) (Histogram of Oriented Gradients (3D)) jest zbliżony do przedstawionego wcześniej deskryptora HOG, lecz agregacja odbywa się w niewielkich komórkach wolumenu przestrzenno-czasowego sąsiadujących z wykrytym punktem zainteresowania, dodatkowo proponowany jest deskryptor HOF (3D) (Histogram of Flow (3D)) obliczany podobnie, ale opierający się na danych przepływu optycznego. Obliczone deskryptory są agregowane za pomocą metody Bag-Of-Words, jednakże agregacja odbywa się oddzielnie dla deskryptorów HOG (3D) i HOF (3D), a dodatkowo stosowany jest czasowo przestrzenny podział punktów zainteresowania, które agregowane są wewnątrz wybranych pod-wolumenów danych w zależności od konfiguracji.

Rozszerzeniem deskryptorów opartych o punkty charakterystyczne w wolumenach przestrzenno-czasowych są deskryptory oparte na trajektoriach wewnątrz wolumenów przestrzenno-czasowych[18]. Autorzy cytowanej pracy próbkują obraz w równych odstępach (stąd nazwa metody to 'gęste trajektorie'). Następnie wszystkie obiecujące punkty (tzn. takie, które nie znajdują się w bardzo jednostajnych obszarach), są śledzone w czasie na podstawie obliczonych wcześniej pól przepływu optycznego. Śledzone punkty tworzą trajektorie, które kodują lokalne cechy ruchu. Procedura wykonywana jest w wielu skalach. Utworzone trajektorie stanowią lokalne deskryptory ruchu. Dodatkowo w sąsiedztwie poszczególnych trajektorii obliczane są znane deskryptory HOG (3D) i HOF (3D) oraz nowe - Motion

[15] Kläser, A., Marszalek M., Schmid C.. „A Spatio-Temporal Descriptor Based on 3D-Gradients," Proceedings of British Machine Vision Conference 2008, 2008.
[16] Willems G., Tuytelaars T., Van Gool L., „An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector,", s. 650-663, 2008.
[17] Laptev I, Marszalek M., Schmid C., Rozenfeld B., „Learning realistic human actions from movies," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, s. 1-8, 2008.
[18] Wang H., Kläser A., Schmid C., Liu C., „Action recognition by dense trajectories," CVPR 2011, Providence, s. 3169-3176, 2011.

Boundary Histogram (MBH) – opisujące gradienty w polu przepływu optycznego, które są szczególnie wrażliwe na obszary obrazu, w których zmienia się charakterystyka ruchu.

Opis ruchu w obrazie za pomocą trajektorii doczekał się też ulepszeń[19]. W cytowanej pracy autorzy zaproponowali kilka modyfikacji oryginalnej metody. Zastosowano m.in. kompensację ruchu w obrazie wywołanego ruchem kamery. Estymacja ruchu kamery opiera się na punktach charakterystycznych SURF i przepływie optycznym (co ciekawe, pomijane są obszary obejmujące osoby – jako potencjalnie najbardziej „zmienne"). Następnie cechy oparte o przepływ optyczny są korygowane na podstawie ruchu kamery, a trajektorie nadmiernie skorelowane z ruchem kamery są eliminowane. Innym ulepszeniem jest zastąpienie klasycznej metody Bag-of-Words nowocześniejszym rozwiązaniem – wektorami Fishera.

METODY AGREGACJI CECH LOKALNYCH

W odróżnieniu od globalnych cech sylwetki, cechy lokalne oparte bardzo często na czasowo-przestrzennych punktach charakterystycznych zwykle posiadają informację uwzględniającą temporalny charakter danych. Z tego powodu modele analizujące sekwencje, takie jak HMM i DTW, mają mniejsze zastosowanie dla tej kategorii metod. Natomiast istotne stają się metody agregujące rozproszoną w sekwencji chmurę punktów charakterystycznych i ich deskryptorów w zwartą informację liczbową o ściśle określonym rozmiarze.

W przypadku rozpoznawania sekwencji najczęściej adaptuje się znane metody agregacji używane również do rozpoznawania obrazów określane zbiorczym mianem Bag-of-Visual-Words (BoVW) lub po prostu Bag-of-Words (BOW). Metody te mają swój początek w narzędziach opisu dokumentów tekstowych, w których dokument podsumowywany był na podstawie zliczania częstości występowania pewnych wyrazów (histogramy częstości). W przypadku zastosowania do obrazów (lub sekwencji), zliczane są częstości występowania pewnych określonych kategorii deskryptorów punktów charakterystycznych wykrytych w obrazie lub sekwencji. Metoda wymaga procesu uczenia, w którym jest określany specyficzny dla danego zastosowania słownik kategorii. Działanie algorytmu BoVW można podsumować w następujący sposób:

Etap uczenia:

1. Deskryptory punktów charakterystycznych zbierane są ze wszystkich sekwencji uczących
2. Realizowany jest algorytm grupowania (klasteryzacji) - np. k-średnich. Obliczone środki klastrów stają się wzorcami w słowniku kategorii deskryptorów

Etap generowania cech:

1. Dla każdego nowego przykładu (obrazu, sekwencji) generowane są punkty charakterystyczne oraz ich deskryptory
2. Obliczone deskryptory przypisywane są do wcześniej określonych kategorii przez porównanie z deskryptorem wzorcowymi

[19] Wang H, and Schmid, C. „Action Recognition with Improved Trajectories," 2013 IEEE International Conference on Computer Vision, Sydney, NSW, s. 3551-3558, 2013.

3. Określane są liczby wystąpień deskryptorów w poszczególnych kategoriach i określany jest histogram wystąpień. Histogram może być wykorzystany jako opis zawartości badanego obrazu lub sekwencji.

Metoda agregacji BoVW okazuje się być bardzo skuteczna w zadaniach rozpoznawania obrazów i wideo, ze względu na dużą zdolność do odwzorowania struktury zawartości obiektów w obrazie lub sekwencji. Natomiast co do zasady metoda zupełnie ignoruje wzajemną lokalizację elementów obrazu (stąd też nazwa – worek słów), co jednak zwykle nie wpływa negatywnie na wynik klasyfikacji.

Oryginalna metoda BoVW ma rozszerzenia, które w istotny sposób wpływają na jakość reprezentacji. Jednym z nich jest tzw. „miękkie" przypisanie do klastrów. W rozwiązaniu tym, ta etapie uczenia przestrzeń deskryptorów jest przybliżana nie przez zbiór rozłącznych klastrów, lecz przez mieszaninę rozkładów Gaussa o różnych parametrach. W takiej sytuacji każdemu nowemu deskryptorowi możemy przypisać stopień „miękkiej" przynależności do poszczególnych rozkładów, zamiast kategorycznej przynależności, bądź nie do jednego z nich. Dalszym rozwinięciem tej koncepcji są wektory Fishera, które do opisu przynależności używają pochodnych rozkładów – czyniąc go jeszcze bardziej szczegółowym[20].

METODY OPARTE O CECHY GENEROWANE PRZEZ GŁĘBOKIE SIECI NEURONOWE

Wraz z upowszechnieniem się głębokich sieci neuronowych, dostępności dużych zbiorów danych oraz narzędzi do uczenia dużych modeli, zmniejszyło się zapotrzebowanie na wektory cech, które były określane na podstawie wiedzy eksperckiej. Natomiast duży nacisk został położony na rozwój cech generowanych automatycznie za pomocą algorytmów uczących sieci neuronowe. W tym zagadnieniu fundamentalną rolę odgrywają sieci neuronowe o charakterze splotowym (Convolutional Neural Networks – CNN). W tego rodzaju sieciach pierwsze etapy przetwarzania polegają na poddaniu obrazu filtracji przy wykorzystaniu predefiniowanego zbioru filtrów (przypominających filtry używane np. w korekcie fotografii). Tego rodzaju filtry były już z powodzeniem wykorzystywane do generowania cech wcześniej, istotną różnicą jest jednak to, że w przypadku CNN parametry tych filtrów nie są ustalane arbitralnie, lecz uczone na podstawie wejściowej bazy danych i zadanego problemu. W klasycznej sieci CNN cechy generowane są począwszy od najwyższego do coraz niższego poziomu szczegółowości. Sieci oparte o cechy splotowe wykazały dużą skuteczność w problemach rozpoznawania obrazów, ale również sekwencji wideo.

Jedną z pierwszych istotnych architektur sieci splotowych do rozpoznawania aktywności w wideo był ConvNet[21]. Autorzy tej sieci najpierw ekstrahują z badanej sekwencji wideo podsekwencję klatek. Następnie jedna z klatek wybierana jest jako reprezentant podsekwencji. Dodatkowo dla wszystkich klatek podsekwencji tworzone są wektory ruchu przy pomocy metody przepływu optycznego i takie dane

[20] Wang H, and Schmid, C. op. cit.

[21] Simonyan K. and Zisserman A., „Two-stream convolutional networks for action recognition in videos," Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14), MIT Press, tom 1, s. 568–576, Cambridge, MA, USA, 2014.

agregowane w czasie tworzą znowu pojedyczny obraz, ale o dużej liczbie kanałów (analizowane są też warianty, np. próbkowanie wektorów ruchu wzdłuż trajektorii oraz próbkowanie w przód i wstecz osi czasu, wykorzystanie transferu wiedzy z innych sieci). Tak utworzone dwa obrazy (obejmujące dane obrazu oraz dane o ruchu) są przetwarzane przez niezależne sieci splotowe, a ich wyniki agregowane. W ramach dłuższych sekwencji stosowane jest proste uśrednianie uzyskanych ocen. Przytoczona praca jest ciekawym przykładem „spłaszczenia" problemu przetwarzania danych przestrzenno-temporalnych do dwóch wymiarów.

Sieci splotowe mogą zostać połączone z sieciami rekurencyjnymi w celu uzyskania większej skuteczności[22]. W cytowanym rozwiązaniu każda klatka sekwencji jest przetwarzana niezależnie (jak obraz) przez sieć splotową oraz warstwy w pełni połączonej sieci neuronowej. Następnie wyniki działania sieci dla poszczególnych klatek są agregowane w czasie za pomocą sieci rekurencyjnej (w tym rozwiązaniu użyto LSTM operującym w przód i wstecz osi czasu).

Carreira i Zisserman proponują analizę sekwencji obrazów za pomocą splotów trójwymiarowych[23]. Filtry realizujące takie sploty działają nie tylko wzdłuż osi x oraz y, ale również wzdłuż osi czasu. W cytowanej pracy przedstawiony jest również sposób na użycie sieci pre-trenowanych na obrazach statycznych jako punktu wyjścia do uczenia sieci ze splotami 3D. Można tam odnaleźć kompleksowe porównanie metod rozpoznawania aktywności w podziale na różne sposoby agregacji danych obrazowych w czasie oraz różne sposoby uwzględnienia danych or ruchu na podstawie pól przepływu optycznego. Przeprowadzone eksperymenty wskazują, że najlepsze wyniki osiągane są przez 2 niezależne sieci działające na danych obrazowych oraz danych przepływu optycznego, których wyniki uzgadniane są na końcu procesu rozpoznania (tzw. architektura Two-Stream 3D-ConvNet).

SPOSOBY AGREGACJI CECH GENEROWANYCH PRZEZ SIECI NEURONOWE W CZASIE

Podobnie, jak w przypadku metod klasycznych, także w przypadku sieci splotowych problem rozpoznawania sekwencji wideo najczęściej wymaga agregacji informacji w osi czasu. Najczęściej wykorzystywane sieci neuronowe o charakterze splotowym koncentrują się na pojedynczej klatce obrazu, pozostawiając interpretację sekwencji narzędziom wyższego poziomu. Najczęściej wykorzystywane metody umożliwiające agregację informacji w osi czasu to:

- sieci neuronowe w pełni połączone, które pozyskują dane z poszczególnych klatek, reorganizują i przetwarzają za pomocą sieci w pełni połączonej typu perceptron. Pewnym problemem tych rozwiązań jest zazwyczaj dosyć duża liczba parametrów takich sieci co utrudnia uczenie
- proste metody agregacji w czasie polegające na zagregowaniu za pomocą prostej operacji (średnia lub maksimum) rezultatów przetwarzania dla pojedynczych klatek. Problemem tego typu metod jest ignorowanie interakcji

[22] Ullah A., Ahmad J., Muhammad K., Sajjad M., Baik S. W., „Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," IEEE Access, tom. 6, s. 1155-1166, 2018.
[23] Carreira J., Zisserman A., „Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, s. 4724-4733, 2017.

między poszczególnymi klatkami – natomiast metoda może być łączona z innymi (np. splotami 3D)
- sploty 3D oraz lokalne agregacje w wolumenach 3D (Max/Avg Pooling 3D)– użycie tych metod pozwala na agregację informacji temporalnych na bardzo niskim poziomie, kosztem jest pewne zwiększenie liczby parametrów sieci. Sieci oparte o sploty 3D zasadniczo uwzględniają tylko niewielki kontekst czasowy wokół analizowanej klatki.
- wykorzystanie sieci rekurencyjnych. Sieci rekurencyjne przetwarzają kolejno dane z kolejnych klatek, uwzględniając zapamiętaną dotychczas historię (zawsze zapamiętują stan, na który wpływają nowe dane – taka charakterystyka czyni je bardzo użytecznymi w analizie sekwencji). Obecnie najczęściej wykorzystywane są sieci stosujące zaawansowane zarządzanie pamięcią (specjalne mechanizmy retencji i czyszczenia pamięci), takie jak LSTM (Long Short-time Memory) lub GRU (Gated Recurrent Unit). Mimo zastosowania zaawansowanych metod, sieci takie mogą mieć problemy w kojarzeniu bardzo odległych zjawisk (np. z początku i końca sekwencji).
- Wykorzystanie sieci typu transformer, które za pomocą mechanizmu atencji posiadają umiejętność kojarzenia nawet bardzo odległych elementów sekwencji czasowej

SIECI NEURONOWE TYPU TRANSFORMER

Ostatnio zyskujące na popularności sieci typu transformer cechują się inną filozofią przetwarzania danych od sieci splotowych. Podczas gdy sieci splotowe koncentrują się na lokalnym sąsiedztwie każdego elementu danych (np. piksela w obrazie), w sieciach typu transformer koncepcja sąsiedztwa praktycznie nie istnieje (trzeba ją wymuszać za pomocą dodatkowych narzędzi), a wszystkie jednostki danych są traktowane w podobny sposób. W sieciach typu transformer analiza danych opiera się na kojarzeniu poszczególnych jednostek (przetworzonych) danych (tokenów) za pomocą mechanizmów korelacji – tworząc tzw. mapę atencji. Mapa ta jest używana do przypisywania wag skojarzonym elementom, które są następnie kombinowane.

Taka filozofia przetwarzania danych eliminuje problemy sieci związane z trudnością kojarzenia odległych od siebie elementów (które były charakterystyczne dla sieci splotowych, ale również dla sieci rekurencyjnych). Natomiast w celu zachowania potencjału sieci w zakresie uwzględnienia również pozycji elementów (co jest jednak istotne w danych obrazowych oraz w sekwencjach czasowych) stosuje się specjalne zabiegi „znakowania pozycji" danych wejściowych (w przeciwnym razie transformer działałby na zasadzie zbliżonej do metod BoVW).

Transformery odniosły duży sukces w przetwarzaniu danych tekstowych, natomiast są też z powodzeniem wykorzystywane analizie obrazów oraz wideo. Prawdopodobnie pierwszym zastosowaniem architektury Transformer dla obrazów był Transformer Wizyjny[24] (Vision Transformer - ViT). W rozwiązaniu wykorzystywana jest architektura sieci transformer oparta tylko o enkoder. Ponieważ co do zasady transformery przyjmują dane sekwencyjne obraz jest

[24] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai, X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N., „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR'21, 2021.

dzielony na komórki, które po transformacji oraz dodaniu znakowania pozycji służą jako tokeny wejściowe. Ostateczna klasa obiektu jest wyznaczana przez w pełni połączoną sieć neuronową.

ViViT[25] stanowi udaną próbę zastosowania sieci transformer do analizy sekwencji obrazów z uwzględnieniem ich specyfiki. Autorzy architektury wskazują na istotny problem w bezpośrednim zastosowaniu architektury ViT, polegający na istotnym zwiększeniu liczby parametrów oraz złożoności obliczeniowej zadania, ze względu na znacznie większy rozmiar przestrzeni wejściowej (wiele obrazów zamiast jednego). Autorzy analizują różne konfiguracje, w których operacja atencji w czasie i przestrzeni nie byłaby prowadzona jednocześnie a rozdzielona (faktoryzowana), co istotnie wpływa na złożoność obliczeniową modelu.

Podobne problemy stoją również u podłoża powstania modelu TimeSformer[26]. Autorzy oprócz prostej faktoryzacji atencji w czasie i przestrzeni, proponują tam kilka innych wariantów faktoryzacji np. przez rzadkie próbkowanie w czasie i przestrzeni obrazu lub próbkowanie tylko wzdłuż osi (X,Y,T).

Rozmiary wolumenów danych, na których działa transformer nie muszą zawsze być stałe, ale mogą zmieniać się w kolejnych etapach przetwarzania. Taka koncepcja została zastosowana w modelu Multiscale Vision Transformer - MViT[27]. W pierwszych etapach przetwarzania stosuje się tu wolumen danych w pełnej rozdzielczości, ale utrzymując małą liczbę kanałów, w kolejnych etapach rozdzielczość się zmniejsza, ale zwiększa się jednocześnie liczbę używanych kanałów. W ten sposób utrzymywana jest podobna złożoność obliczeniowa na każdym etapie przetwarzania danych. Proponowana architektura została rozwinięta w ramach modelu MViTv2[28].

Olbrzymie rozmiary danych wideo wykorzystywane do uczenia sieci są problemem w przypadku korzystania z sieci typu transformer, zatem często autorzy prac starają się ograniczyć liczbę używanych tokenów, od czego zależy złożoność obliczeniowa. Przykładowo TokenLearner[29] jest specjalną warstwą atencyjną (atencji przestrzennej), na podstawie której realizowana jest redukcja liczby tokenów wykorzystywanych przez kolejne warstwy sieci. Proces ten może być odwrócony za pomocą symetrycznej warstwy TokenFuser. Zmniejszenie liczby tokenów daje istotne przyspieszenie obliczeń oraz wpływa pozytywnie na skuteczność.

Wyzwanie polegające na doborze optymalnego zestawu tokenów, pod względem rozmiaru i liczby, jest również inspiracją dla architektury Multiview

[25] Anurag A., Dehghani M., Heigold G., Sun Ch., Lucic M., Schmid C., „ViViT: A Video Vision Transformer.", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), s. 6816-6826, 2021.

[26] Bertasius G., Wang H., Torresani L. Is Space-Time Attention All You Need for Video Understanding?, 2021.

[27] Fan H., Xiong B., Mangalam K., Li Y., Yan Z., Malik J., Feichtenhofer C., „Multiscale Vision Transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), s. 6824–35, 2021.

[28] Li Y., Wu C.-Y., Fan H., Mangalam K., Xiong B., Malik, J., Feichtenhofer, C., „Mvitv2: Improved multiscale vision transformers for classification and detection," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), s. 4794–4804. 2022.

[29] Ryoo M., Piergiovanni A., Arnab A., Dehghani M., Angelova A., Ranzato M., „TokenLearner: Adaptive Space-Time Tokenization for Videos", Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS'21), Curran Associates Inc., art. numer. 979, s. 12786-12797, 2021.

Transformer[30]. Autorzy proponują elastyczne podejście, w którym różne kombinacje zestawów tokenów (zwanych widokami) działają równolegle, a ich wyniki są integrowane za pomocą metod fuzji widoków. W zamyśle autorów tokeny obejmujące duże interwały czasowe lepiej reprezentują cechy globalne nagrania (takie jak tło sceny), podczas gdy tokeny o krótszym czasie trwania krótkotrwałe czynności (np. gestykulację).

Coraz częściej autorzy sieci starają się wykorzystać synergie pomiędzy różnymi zbiorami danych oraz zadaniami uczenia maszynowego w celu poprawy wyników sieci[31]. W cytowanej pracy przedstawione jest rozszerzenie sprawdzonego modelu TimeSformer, w którym stosowane są zaawansowane metody transferu wiedzy pomiędzy modelami i zbiorami danych. W pierwszej kolejności model jest wstępnie trenowany na bazie obrazów statycznych (przy usuniętych mechanizmach atencji czasowej), a dopiero potem również na danych wideo. Dodatkowo stosowany jest mechanizm uczenia wielozadaniowego (Multi-task Learning), w którym uczenie następuje na kilku różnych zbiorach danych wideo oraz danych statycznych jednocześnie (stosując nieco odmienne warstwy klasyfikacyjne). Dzięki temu podejściu system uczy się nie tylko cech dostępnych w specyficznym zbiorze danych, ale korzysta też cech dopasowanych do innych zbiorów danych, co zwiększa jego elastyczność.

Nowatorska architektura UniFormer[32] wykorzystuje spójny formalizm modułów atencyjnych do reprezentowania zarówno mechanizmów korzystających z lokalnego przestrzenno/temporalnego kontekstu (ideowo zbliżonych do splotów 3D) oraz mechanizmów atencji globalnej (ideowo zbliżonych do klasycznych modułów Multi-Head-Self-Attention). Pierwsze zajmują się lokalną agregacją cech, natomiast drugie modelują bardziej odległe relacje czasowo-przestrzenne. Druga wersja architektury UniFormerV2[33] oprócz używania podobnych mechanizmów potrafi wykorzystać pre-trenowane wagi modelu ViT stosując faktoryzację przetwarzania czasowego i przestrzennego w sposób zbliżony do zaproponowanego w ViViT.

Architektura InternVideo[34] odwołuje się do zyskującego na popularności paradygmatu uczenia samo-nadzorującego się (Self-Supervised Learning) oraz modeli bazowych (Foundation Models). W proponowanym rozwiązaniu wykorzystane są dwa zaawansowane mechanizmy wstępnego uczenia sieci transformer na dużych zbiorach danych – uczenie autoenkodera przez maskowanie wejścia oraz multimodalne uczenie kontrastowe. Wymiana wiedzy pomiędzy tak nauczonymi sieciami następuje przy użyciu mechanizmu atencji skrośnej. Tak przygotowany model bazowy może być następnie używany w różnych zadaniach, m.in. do rozpoznawania aktywności w wideo.

---

[30] Yan S., Xiong X., Arnab A., Lu Z., Zhang M.,Sun C., Schmid C., „Multiview Transformers for Video Recognition," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, s. 3323-3333, 2022.

[31] Zhang B., Yu J,, Fifty C., Han W., Dai A.M., Pang R., & Sha F., „Co-training Transformer with Videos and Images Improves Action Recognition", 2021, url: https://arxiv.org/abs/2112.07175

[32] Li K. & Wang Y., Gao P., Song G, & Liu Y., Li H., Qiao Y. „UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning", 2022.

[33] Li K., Wang Y., He Y., Li Y., Wang Y., Wang L., Qiao Y., UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer, 2022, https://arxiv.org/abs/2211.09552.

[34] Wang Y., Li K., Li Y., He Y., Huang B., Zhao Z., Zhang H, Xu J., Liu Y., Wang Z., Xing S. & Chen G. & Pan J., Yu J., Wang Y., Wang L., Qiao Y, InternVideo: General Video Foundation Models via Generative and Discriminative Learning, 2022, https://arxiv.org/abs/2212.03191.

Uczenie wielomodalne, wielozadaniowe i wykorzystanie synergii między modalnościami i zadaniami umożliwia zwiększenie skuteczności rozpoznawania[35]. Autorzy cytowanej pracy stworzyli system oparty o sieć transformer umożliwiający przetwarzane danych z aż 12 modalności (wliczając obraz, dźwięk, audio, tekst) oraz kilka zadań uczenia maszynowego. Model uczony jest zawsze na parze modalności (oraz parze zadań) i posiada zunifikowane wagi dla obu modalności za wyjątkiem specyficznych dla modalności tokenizerów. Oprócz współdzielenia wag wykorzystywany jest moduł atencji skrośnej w celu wymiany informacji między modalnościami. Generowane przez model uniwersalne cechy sprawdzają się bardzo dobrze w wielu zadaniach m.in. w rozpoznawaniu aktywności.

PODSUMOWANIE

Zrealizowany przegląd najciekawszych metod rozpoznawania aktywności wskazuje na potencjalne kierunki rozwoju algorytmów ekstrakcji cech i ich dalszego przetwarzania. Wydaje się, że cechy oparte na wiedzy eksperckiej wypierane są przez zaawansowane metody analizy obrazu oparte o sieci neuronowe, a ostatnio o mechanizmy atencji. Rozwiązania tego typu, oparte np. o sieć transformer, były z powodzeniem używane do rozwiązywania uniwersalnego problemu rozpoznawania aktywności, w oparciu o wciąż rosnącą moc obliczeniową i dostępność coraz obszerniejszych zbiorów danych i są wdrażane do specyficznych problemów, jak rozpoznawanie materiałów pornograficznych, czy zachowań agresywnych.

Patrząc na szerszy kontekst, w najnowszych rozwiązaniach widać tendencję do wykorzystywania wielu zbiorów danych, wielu modalności i wielu zadań uczenia maszynowego równocześnie, jak również strategii uczenia samo-nadzorującego się w celu wypracowania wspólnej (i w dużym stopniu uniwersalnej) reprezentacji danych wejściowych umożliwiającej lepsze zrozumienie poszczególnych podproblemów przez sieć neuronową.

Bibliografia

Achard C., & Qu X. & Mokhber A. & Milgram M. „A novel approach for recognition of human actions with semi-global features". Mach. Vis. Appl.. tom. 19. s. 27-34., 2008, doi:10.1007/s00138-007-0074-2.

Adamiok F., Wilkowski A., Comparison of deep learning approaches to violence detection in videos, Progress in Polish Artificial Intelligence Research 5. Proceedings of the 5th Polish Conference on Artifcial Intelligence (PP-RAI'2024) 18–20.04.2024, Warsaw, Poland, ed. Mańdziuk Jacek, Żychowski Adam, Małkiński Mikołaj (red.), Politechnika Warszawska, s.249-256, 2024, ISBN 978-83-8156-696-4

Anurag A., Dehghani M., Heigold G., Sun Ch., Lucic M., Schmid C., „ViViT: A Video Vision Transformer.", 2021 IEEE/CVF International Conference on Computer Vision (ICCV), s. 6816-6826, 2021

Bertasius G., Wang H., Torresani L. Is Space-Time Attention All You Need for Video Understanding?, 2021, doi:10.48550/arXiv.2102.05095.

Bobick A. F., Davis J. W., „The recognition of human movement using temporal templates," w IEEE Transactions on Pattern Analysis and Machine Intelligence, tom: 23(3), s. 257-267, 2001, doi: 10.1109/34.910878.

[35] Srivastava S., Sharma G., „OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), s. 27402-27414, Seattle, WA, USA, 2024.

Carreira J., Zisserman A., „Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, s. 4724-4733, 2017, doi: 10.1109/CVPR.2017.502.

Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai, X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N., „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", ICLR'21, 2021

Efros A.A., Berg A.C., Mori G. and Malik J., „Recognizing action at a distance," Proceedings Ninth IEEE International Conference on Computer Vision, Nice, France, tom. 2, s. 726-733 2003, doi: 10.1109/ICCV.2003.1238420.

Fan H., Xiong B., Mangalam K., Li Y., Yan Z., Malik J., Feichtenhofer C., „Multiscale Vision Transformers," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), s. 6824–35, 2021.

Kläser, A., Marszalek M., Schmid C.. „A Spatio-Temporal Descriptor Based on 3D-Gradients," Proceedings of British Machine Vision Conference 2008, 2008, doi: 10.5244/C.22.99

Laptev I., Lindeberg, T. Local Descriptors for Spatio-temporal Recognition. ECCV, tom 3667. s. 91-103, 2004, doi:10.1007/11676959_8.

Laptev I, Marszalek M., Schmid C., Rozenfeld B., „Learning realistic human actions from movies," 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, s. 1-8, 2008, , doi:10.1109/CVPR.2008.4587756.

Li Y., Wu C.-Y., Fan H., Mangalam K., Xiong B., Malik, J., Feichtenhofer, C., „Mvitv2: Improved multiscale vision transformers for classification and detection," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), s. 4794–4804. 2022. doi:10.1109/CVPR52688.2022.00476

Li K. & Wang Y., Gao P., Song G, & Liu Y., Li H., Qiao Y. „UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning", 2022, 10.48550/arXiv.2201.04676.

Li K., Wang Y., He Y., Li Y., Wang Y., Wang L., Qiao Y., UniFormerV2: Spatiotemporal Learning by Arming Image ViTs with Video UniFormer, 2022, https://arxiv.org/abs/2211.09552

Lu W.-L., Little J. J., „Simultaneous Tracking and Action Recognition using the PCA-HOG Descriptor," The 3rd Canadian Conference on Computer and Robot Vision (CRV'06), Quebec, Canada, s. 6-6, 2006, doi: 10.1109/CRV.2006.66.

Myers C., Rabiner L., Rosenberg A., „Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," Acoustics, Speech, and Signal Processing, IEEE Transactions on, tom 28(6), s. 623–635, 1980.

Niewiadomska-Szynkiewicz E., Różycka M., Staciwa K., Nyczka K., „System wspomagający wykrywanie treści wizualnych i tekstowych zagrażających bezpieczeństwu dzieci w cyberprzestrzeni." Cybersecurity and Law 2023, nr 2(10), s. 202-220, 2023

Oikonomopoulos A., Patras I., Pantic M., „Spatiotemporal salient points for visual recognition of human actions," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), tom 36(3), s. 710-719, 2006, doi: 10.1109/TSMCB.2005.861864.

Rabiner L.R., „A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, tom 77(2), s. 257-286, 1989, doi: 10.1109/5.18626.

Ryoo M., Piergiovanni A., Arnab A., Dehghani M., Angelova A., Ranzato M., „TokenLearner: Adaptive Space-Time Tokenization for Videos", Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS'21), Curran Associates Inc., art. numer. 979, s. 12786-12797, 2021

Scovanner P., Ali, S., Shah, M., „A 3-dimensional SIFT descriptor and its application to action recognition," Proceedings of the ACM International Multimedia Conference and Exhibition. S. 357-360, 2007, doi:10.1145/1291233.1291311

Simonyan K. and Zisserman A., „Two-stream convolutional networks for action recognition in videos," Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS'14), MIT Press, tom 1, s. 568–576, Cambridge, MA, USA, 2014

Srivastava S., Sharma G., „OmniVec2 - A Novel Transformer Based Network for Large Scale Multimodal and Multitask Learning," 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), s. 27402-27414, Seattle, WA, USA, 2024, doi: 10.1109/CVPR52733.2024.02588.

Sundaresan A., RoyChowdhury A., Chellappa R., „A hidden Markov model based framework for recognition of humans from gait sequences,", Proceedings 2003 International Conference on

Image Processing (Cat. No.03CH37429), Barcelona, Spain, , s. II-93, 2003, doi: 10.1109/ICIP.2003.1246624.
Ullah A., Ahmad J., Muhammad K., Sajjad M., Baik S. W., „Action Recognition in Video Sequences using Deep Bi-Directional LSTM With CNN Features," IEEE Access, tom. 6, s. 1155-1166, 2018, doi: 10.1109/ACCESS.2017.2778011.
Veeraraghavan A., Chowdhury A. R., Chellappa R., „Role of shape and kinematics in human movement analysis", Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004, pp. I–730, Washington, DC, USA, 2004
Wang H., Kläser A., Schmid C., Liu C., „Action recognition by dense trajectories," CVPR 2011, Providence, s. 3169-3176, 2011, doi: 10.1109/CVPR.2011.5995407.
Wang H, and Schmid, C. „Action Recognition with Improved Trajectories," 2013 IEEE International Conference on Computer Vision, Sydney, NSW, s. 3551-3558, 2013, doi: 10.1109/ICCV.2013.441.
Willems G., Tuytelaars T., Van Gool L., „An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector,", s. 650-663, 2008 doi:10.1007/978-3-540-88688-4_48
Wang Y., Li K., Li Y., He Y., Huang B., Zhao Z., Zhang H, Xu J., Liu Y., Wang Z., Xing S. & Chen G. & Pan J., Yu J., Wang Y., Wang L., Qiao Y, InternVideo: General Video Foundation Models via Generative and Discriminative Learning, 2022, doi:10.48550/arXiv.2212.03191, https://arxiv.org/abs/2212.03191,
Yan S.. Xiong X., Arnab A., Lu Z., Zhang M.,Sun C., Schmid C., „Multiview Transformers for Video Recognition," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, s. 3323-3333, 2022, doi: 10.1109/CVPR52688.2022.00333
Zhang S., Wei Z., Nie J., Huang L., Wang S. & Li Z. „A Review on Human Activity Recognition Using Vision-Based Method. Journal of Healthcare Engineering". s. 1-31, 2017 doi: 10.1155/2017/3090343.
Zhang B., Yu J,, Fifty C., Han W., Dai A.M., Pang R., & Sha F., „Co-training Transformer with Videos and Images Improves Action Recognition", 2021, url: https://arxiv.org/abs/2112.07175

METHODS FOR ACTIVITY RECOGNITION IN VIDEO SEQUENCES USING LOW-LEVEL IMAGE FEATURES

Abstract

Issues related to the problem of recognizing activity in video footage are at the center of many public security services' concerns. Adequate recognition systems can enhance citizen security by detecting dangerous, violent and illegal behavior in video surveillance or by detecting and filtering unwanted and illegal videos available on the Internet, including pornographic or CSAM materials. Data extracted from video may be low-level features (e.g., color, motion) or high-level features (e.g., detected joint positions of human silhouettes). The article undertakes the task of presenting computer solutions for video classification with a special focus on the use of low-level features. Classical methods and the latest methods using deep learning are presented.

Keywords: human activity recognition, video processing, neural networks.