FKT SCIENCE

WWW.FKT-ONLINE.DE

7|2019

FACHZEITSCHRIFT FÜR
FERNSEHEN, FILM UND
ELEKTRONISCHE MEDIEN

RTM
RUNDFUNKTECHNISCHE
MITTEILUNGEN DES IRT

AUTOMATISCHE KAMERAPFAD-GENERIERUNG AUS 360°-VIDEO MIT DEEP-LEARNING-METHODEN

HANNES FASSOLD

Deep Learning ist eine disruptive Technologie und sehr wertvoll für die Extraktion semantischer Informationen aus Videoinhalten. In diesem Beitrag beschreiben wir einen neuartigen Algorithmus zur automatischen Erzeugung eines konventionellen Videos (für den passiven Konsum, ohne Interaktion) aus einem 360°-Video basierend auf semantischer Information, die mit Deep-Learning-Methoden gewonnen wurde. Des Weiteren wird die Deep-Learning-basierte Methode zur automatischen Extraktion der Szenenobjekte (Personen, Tiere, Autos usw.) beschrieben, die im Algorithmus verwendet wird.

Deep learning is a disruptive technology and particularly valuable for extracting semantic information from video content. In this contribution, we are describing a new algorithm which allows automatic generation of a conventional video (for passive consumption, without interaction) from a 360° video. This is based on semantic information obtained using deep learning methods. Furthermore, the deep learning based method for automatic extraction of scene objects (people, animals, cars, etc.) used in the algorithm is described.

Einführung

360°-Video ist in letzter Zeit sehr populär geworden, weil es dem Betrachter ermöglicht den Inhalt in einer sehr unmittelbaren und eindringlichen Art zu erleben. Es wird typischerweise interaktiv konsumiert, indem der Betrachter aktiv navigiert, um jenen Ausschnitt des Videos zu wählen, den er gerade betrachten möchten. Allerdings verfügen nicht alle Geräte über Möglichkeiten zur interaktiven Navigation. Ältere Fernsehgeräte aus der Ära vor Smart TV stellen natürlich keinen interaktiven Player für 360°-Video zur Verfügung. Weiterhin kann es selbst auf Geräten mit Unterstützung für interaktive Navigation sein, dass ein Betrachter den passiven Konsum des 360°-Videos bevorzugt, ohne selbst einzugreifen (Lean-Back-Verhalten). Wir schlagen deshalb die automatische Generierung eines Kamerapfades vor, der den Benutzer durch das Video führt. Ziel ist es daher, automatisch einen visuell interessanten Kamerapfad aus einem 360°-Video zu berechnen, um ein passives Erlebnis wie bei der Betrachtung eines „traditionellen" Videos zu ermöglichen. Es soll also aus dem 360°-Video automatisch ein konventionelles Video generiert werden, indem ein Kamerapfad berechnet wird, der jeweils die momentan interessanteste Region zeigt. Diese Regionen korrespondieren oft mit Personen, die bestimmte Handlungen ausführen, z. B. in einer Konzertszene sind die Bandmitglieder auf der Bühne natürlich am interessantesten.

Um automatisch einen ansprechenden und visuell interessanten Kamerapfad zu generieren, ist eine zeitlich-örtliche Beschreibung des 360°-Videos auf einer hohen semantischen Ebene notwendig. Diese semantische Information wird typischerweise durch das Fusionieren der Resultate von grundlegenden Bildverarbeitungsalgorithmen (z. B. für Objektdetektion oder Bildsegmentierung) generiert. Mit Deep Learning – ermöglicht durch die immense Rechenleistung von Grafikkarten (GPUs) – hat sich die Qualität und Laufzeit dieser grundlegenden Bildverarbeitungsalgorithmen dramatisch verbessert, was den Einsatz dieser Methoden in der Praxis erlaubt. Das von uns entwickelten Verfahren zur automatischen Kamerapfad-Generierung stützt sich hauptsächlich auf die automatische Extraktion von bekannten Objekten (wie Personen, Tieren und Autos), die in der Szene vorkommen. Dies semantische Information über die Szenenobjekte erlaubt uns die Berechnung eines Saliency-Maßes, das für jedes Objekt schätzt, wie interessant es für einen Betrachter ist. Danach wird der Kamerapfad durch Tracken des interessantesten Objekts ermittelt. In den folgenden Abschnitten werden die Hauptelemente dieses Verfahrens näher beschrieben.

Automatische Extraktion der Szenenobjekte mit Deep Learning

Die Extraktion von Szenenobjekten befasst sich mit der automatischen Erkennung und Verfolgung aller relevanten Objekte, die in der Szene vorkommen. Die daraus gewonnene semantische Information ist entscheidend für ein gutes Funktionieren des Verfahrens zur automatischen Kamerapfad-Generierung. Im Folgenden beschreiben wir kurz unseren Algorithmus zur automatischen Extraktion von Szenenobjekten. Die Hauptkomponenten unseres Algorithmus sind ein Deep-Learning-basierter Objektdetektor zur Erkennung der Objekte in der Szene, kombiniert mit einem Optical Flow-(Bewegungsfeld-)Algorithmus zum Tracken der Objekte von einem Einzelbild auf das nächste. Für die Detektion der Objekte verwenden wir den YoloV3 Algorithmus [1]. Er ist in der Lage, 80 Objektklassen, die häufig in Videos vorkommen, zu erkennen, beispielsweise Personen, Haustiere (Hunde, Katzen) und Fahrzeuge (Fahrrad, Motorrad, Auto). Weiterhin wird der hochqualitative TV-L1 Optical Flow-Algorithmus [2] zur Berechnung des dichten (pixel-weisen) Bewegungsfeldes zwischen zwei zeitlich aufeinanderfolgenden Einzelbildern eingesetzt. Beide Komponenten laufen vollständig auf der Grafikkarte (GPU).

Für jedes Einzelbild des Videos ist nun der Workflow des Algorithmus wie folgt. Zuerst wird das dichte Bewegungsfeld zwischen dem aktuellen und dem nächsten Einzelbild mit dem Optical Flow Verfahren berechnet. Weiters wird der Objektdetektor , der uns eine Liste von detektierten Objekten (und ihrer Position) liefert, auf dem nächsten Einzelbild aus-

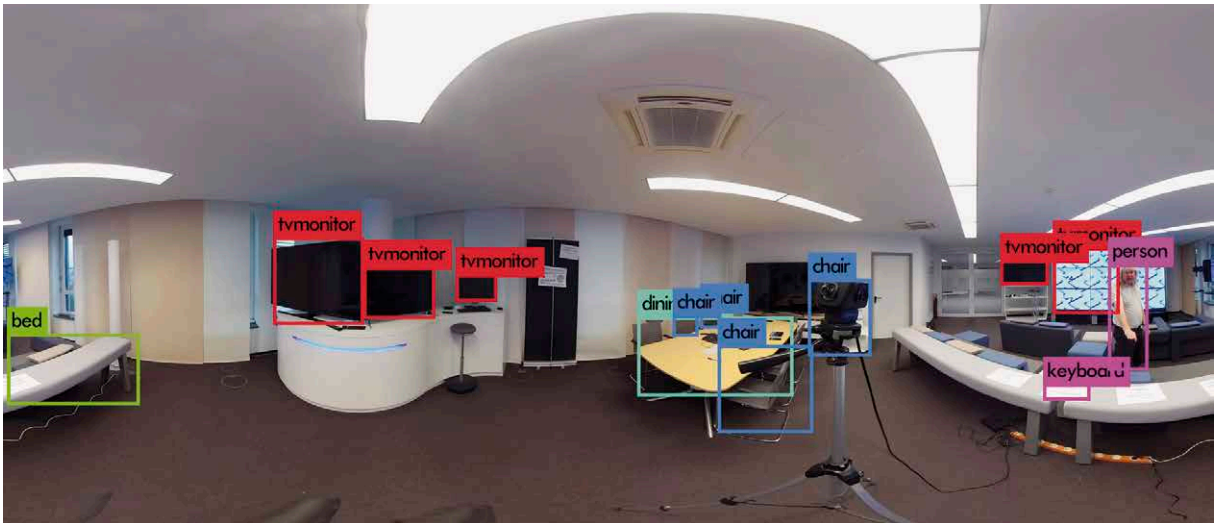

Bild 1. Visualisiertes Resultat der Extraktion von Szenenobjekten für ein Einzelbild eines 360°-Videos
Quelle: der Autor

geführt. Für jedes schon vorhandene Szenenobjekt wird nun mithilfe des Bewegungsfelds die Position dieses Objekts im nächsten Einzelbild vorhergesagt (Tracking). Ein global optimales Matching zwischen den detektierten Objekten und den getrackten Szenenobjekten wird nun durchgeführt. Alle getrackten Szenenobjekte, für die das Matching nicht erfolgreich war (z. B. weil sie von anderen Objekten verdeckt werden), werden als verloren betrachtet und dementsprechend von der Liste der Szenenobjekte entfernt. Alle neu detektierten Objekte, die nicht gematcht werden konnten, werden dagegen zur Liste der Szenenobjekte hinzugefügt. Ein typisches Resultat des Algorithmus für ein Einzelbild aus einem 360°-Video ist in Bild 1 visualisiert. Man sieht, dass der Algorithmus in der Lage ist, die relevanten Objekte in der Szene (wie die Person oder die Monitore) zu erkennen, allerdings treten auch ein paar Falschdetektionen auf (wie das Bett). Die equirektanguläre Projektion des Bildes scheint den Algorithmus nicht wirklich negativ zu beeinflussen.

Automatische Kamerapfad-Generierung

Der Prototyp für die Generierung des automatische Kamerapfads aus einem 360°-Video arbeitet iterativ, d. h. ein Shot (eine Einstellung) wird nach der anderen abgearbeitet. Zunächst wird eine Länge für den nächsten Shot bestimmt. Die Länge des Shots variiert zufällig um einen Basiswert, der vom Benutzer vorgegeben wird, und typischerweise im Bereich von zwei bis zwölf Sekunden liegt. Danach wird der Algorithmus für die automatische Extraktion der Szenenobjekte aufgerufen (siehe voriger Abschnitt). Alle Szenenobjekte, die nicht über den ganzen Shot getrackt werden konnten, werden verworfen, da die Position eines Szenenobjekts in jedem Einzelbild des Shots benötigt wird. Im nächsten Schritt wird eine sogenannte Visited Map berechnet, die es uns erlaubt, den Kamerapfad für den aktuellen Shot in Bereiche des 360°-Videos zu lenken, die in den vorherigen Shots nicht sichtbar waren. Die Visited Map für einen Shot wird berechnet, indem über die Viewports in den vorherigen Shots iteriert wird. Jeder Viewport wird nun in ein gemeinsames Bild projiziert (in equirektangulärer Projektion) und alle projizierten Viewports werden gewichtet aufsummiert. In Bild 2 ist die Visited Map für drei aufeinanderfolgende Shots visualisiert. Die dunklen Bereiche gehören zu Regionen des 360°-Videos, die in letzter Zeit nicht für das Ergebnisvideo genutzt wurden, helle Bereiche waren dagegen im Ergebnisvideo enthalten (also Teil des automatischen Kamerapfads). Für jedes Szenenobjekt werden nun einige Maße berechnet, aus denen das Saliency Maß (Interessantheitsmaß) für dieses Objekt berechnet wird. Für jedes Objekt wird die durchschnittliche Position (x/y) und Dimension (Breite/Höhe des Rechtecks um das Objekt) berechnet, wobei der Durchschnitt über den aktuellen Shot berechnet wird. Wir berechnen die durchschnittliche Stärke der Bewegung des Objekts im aktuellen Shot in gleicher Weise. Zusätzlich wird ein sogenanntes Isoliertheitsmaß berechnet, das uns eine Hinweis gibt wie isoliert das Objekt ist, bezogen auf andere Objekte derselben Objektklasse im gleichen Bild. Als letztes Maß wird ein Visited Maß für das Objekt berechnet aus der durchschnittlichen Objektposition und der Visited Map. Eine Visualisierung der berechneten Maße für jedes Objekt ist in Bild 3 zu sehen.

Basierend auf diesen Maßen wird nun ein Saliency Maß für jedes Objekt berechnet. Dieses Maß liefert uns eine Schätzung dafür, wie interessant ein Objekte für einen Betrachter ist. Das Saliency Maß für ein Objekt basiert auf den folgenden Merkmalen:

- Objektklasse: Natürlicherweise sind Personen typischerweise interessanter als andere Objektklassen (Tiere, Fahrzeuge usw.). Demzufolge wird das Saliency Maß für Objekte vom Typ „Person" erhöht.
- Objektgröße: Das Saliency Maß für das Objekt wird von der durchschnittlichen Objektgröße beeinflusst, sodass große Objekte ein höheres Maß erreichen. Dieses Prinzip ist auch bekannt als die „Hitchcock-Regel" und wird sehr gerne in Filmen eingesetzt.

Bild 2: Visualisierung der Visited Map für drei aufeinanderfolgende Shots
Quelle: der Autor.

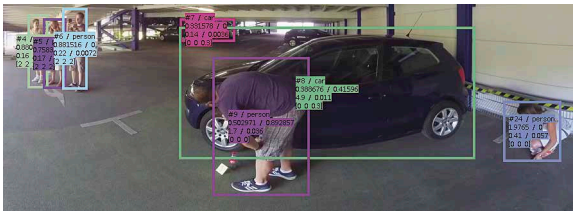

Bild 3: Visualisierung der berechneten Maße für jedes Szenenobjekt
Quelle: der Autor.

- Objektbewegung: Bewegte Objekte sind normalerweise interessanter als statische, weil sie wahrscheinlich eine bestimmte Handlung ausführen. Das Saliency Maß wird daher für bewegte Objekte entsprechend erhöht.
- Isoliertheit: Wir nehmen an, dass isolierte Objekte (wie die Musiker in einem Musikkonzert) wichtiger sind als Objekte derselben Klasse, die sehr eng benachbart sind (wie z. B. Personen im Publikum). Daher wird das Saliency Maß für Objekte mit weniger Nachbarn erhöht.
- Visited Maß: Wir setzen das Visited Maß ein um das Saliency Maß für Objekte, die bereits in den vorherigen Shots zu sehen waren, zu verringern.

Nachdem das Saliency Maß für jedes Objekt berechnet wurde, wird das Objekt mit dem höchsten Saliency Maß als Fokus-Objekt gesetzt. Der automatische Kamerapfad für den aktuellen Shot wird nun generiert, indem das Fokus-Objekt über den ganzen Shot getrackt wird. Bevor die Viewportinformation für jedes Bild des Shots berechnet wird, wird der Pfad des Objekt zeitlich geglättet, um eine glatte Kamerabewegung zu erzeugen. Das Zentrum des Viewports wird auf das Zentrum des Fokusobjekts gesetzt, und der horizontale Öffnungswinkel auf 75°. Bild 4 zeigt das Resultat der automatischen Kamerapfad-Generierung für ein 360°-Video eines Musikkonzerts. Die Basislänge für einen Shot wurde auf drei Sekunden gesetzt, um ein schnell geschnittenes Video zu generieren, das der Dynamik des Inhalts gerecht wird.

Zusammenfassung und Ausblick

Obwohl der aktuelle Prototyp zur automatischen Kamerapfad-Generierung aus 360°-Video brauchbare Ergebnisse liefert, gibt es definitiv Raum für Verbesserungen. Erstens sollte mehr semantische Information aus dem Video extrahiert werden, um den Inhalt besser zu verstehen. Speziell die Information aus der Erkennung von Gesichtern, Emotionen und Handlungen wäre sehr wertvoll. Des Weiteren sollten mehr kinematographische Techniken [3] hinzugefügt werden (wie zum Beispiel Close-Up Shots auf des Gesicht einer wichtigen Person im Video), um das generierte Video abwechslungsreicher zu machen. Wenn Information über die persönlichen Vorlieben eines Betrachter verfügbar ist (z. B. aus Social-Media-Profilen), könnte diese außerdem benutzt werden, um den generierten Kamerapfad besser auf die Vorlieben des Betrachters abzustimmen.

Die Deep-Learning-basierte Methode zur automatischen Extraktion der Szenenobjekte könnte auch für andere Anwendungsgebiete sehr nützlich sein, wie zum Beispiel der semi-automatischen Annotation [4] des Videoinhalts.

Die beschriebenen Forschungs- und Entwicklungstätigkeiten wurden im Rahmen des Forschungsprojekts „Hyper360" (Enriching 360 media with 3D storytelling and personalisation elements) [5] von der Europäischen Union gefördert.

Bild: H. Fassold

HANNES FASSOLD

ist wissenschaftlicher Mitarbeiter im Smart Media Solutions Team bei JOANNEUM RESEARCH – DIGITAL.
www.joanneum.at

Bild 4. Resultat der automatischen Kamerapfad-Generierung für ein 360°-Musikvideo (zu sehen in der ersten Zeile). Jede Zeile zeigt zwei Bilder aus dem generierten Shot (vier aufeinanderfolgende Shots insgesamt).
Quelle: der Autor

Referenzen
[1] Joseph Redmon, Ali Farhadi (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767
[2] Manuel Werlberger, Werner Trobin, Thomas Pock, Andreas Wedel, Daniel Cremers, Horst Bischof (2009). Anisotropic Huber-L1 Optical Flow. In Proceedings of the British Machine Vision Conference (BMVC). London, UK.
[3] Rene Kaiser (2018). Towards Applying the Virtual Director Concept to 360 Degree Video Content, ACM International Conference on Interactive Experiences for TV and Online Video (TVX '18), 2018, Seoul, South Korea.
[4] Hannes Fassold, Barnabas Takacs (2019). Towards Automatic Cinematography and Annotation for 360 Video, ACM International Conference on Interactive Experiences for TV and Online Video (TVX '19), 2019, Salford (Manchester), United Kingdom.
[5] http://www.hyper360.eu/, Horizon 2020 Programm, grant 761934