Automatische Kamerasteuerung bei Vortragsaufzeichnungen

Max Frohberg, Raphael Zender, Ulrike Lucke

Universität Potsdam, Institut für Informatik,
Lehrstuhl für Komplexe Multimediale Anwendungsarchitekturen
A.-Bebel-Str. 89, 14482 Potsdam
vorname.nachname@uni-potsdam.de

Abstract: Der Beitrag stellt eine kombinierte Hardware-Software-Lösung vor, mit der bei Vortragsaufzeichnungen eine automatisierte Erfassung des Sprechers (Dozent oder Fragender aus dem Publikum) erfolgt und in entsprechende Steuersignale für die Kamera umgesetzt wird. Anders als in früheren Entwicklungen greift das System nicht auf spezielle Geräte oder Sensoren am Nutzer zurück, sondern ist minimal-invasiv.

1 Motivation

Vortragsaufzeichnungen und -übertragungen (sog. E-Lectures) erfreuen sich seit mehreren Jahren großer Popularität [TLH09][We+12]. Sie können helfen, überfüllte Hörsäle zu entlasten, geographische Entfernungen zu überbrücken oder Terminkollisionen zu begegnen. Studierende mit kognitiven oder körperlichen Behinderungen, mit sprachlichen Schwierigkeiten oder mit anderen Verpflichtungen können das Lerntempo an ihre persönlichen Bedürfnisse anpassen. Zudem sind Produktion und Distribution inzwischen so einfach und preiswert, dass bereits mehrere Hochschulen die E-Lecture neben der Präsenzveranstaltung zu einem Standardangebot gemacht haben.

Herausforderungen für die Informatik bezogen sich in der Vergangenheit vornehmlich auf den Umgang mit der fertigen Aufzeichnung. Jedoch betrifft eine Reihe von Problemen, die durch Informatik-Systeme gelöst werden können, bereits den Vorgang der Aufzeichnung selbst. Dozenten bewegen sich oft frei auf dem Podium, was entweder einen Kameramann oder einen großen Aufnahmebereich erfordert. Letzteres vergrößert aber zugleich die (physische & soziale) Distanz zum Betrachter. Auch greifen viele Dozenten nicht auf die Zeige-Werkzeuge der Software zurück, sondern gestikulieren wie gewohnt mit ihren Händen. Kameras können das bedingt erfassen, doch Zeigegesten gehen verloren, sofern nicht die gesamte Leinwand mit aufgezeichnet wird. Und auch Fragen aus dem Publikum sind schwer aufzuzeichnen, denn der jeweilige Sprecher muss lokalisiert und in Bild und Ton erfasst werden. Ein wichtiges Akzeptanzkriterium ist dabei, dass die eingesetzte Lösung so wenig wie irgend möglich in die reguläre

Lehrveranstaltung eingreift. Spezielle Geräte oder Software für Dozent oder Publikum stellen eine Einstiegshürde dar, können die kognitive Belastung im Lehr-/Lernprozess vergrößern und sind daher zu vermeiden.

Diesen Problemen widmet sich die hier vorgestellte Lösung. Zeigegesten [Luc12] liegen dabei außerhalb des Fokus dieses Artikels, doch eine automatisierte Lösung zur Fokussierung der Kamera auf die Position des Sprechers (egal ob Dozent oder Publikum) zur Steuerung der Kamera (schwenken und zoomen) wird nachfolgend präsentiert.

2 Nicht-invasive Ortung der Sprecher

Aus dem Spektrum verfügbarer Ortungsmechanismen wurden auditive und optische Verfahren umgesetzt. Auditive Verfahren eignen sich aufgrund der Schallausbreitungseigenschaften vorwiegend für das Publikum während optische Verfahren zwar eine genauere aber aufgrund des beschränkten Sichtfeldes nur für den Dozenten geeignete Erfassung ermöglichen.

2.1 Auditive Ortung

Das System setzt die Implementierung eines Beamforming-Algorithmus [BK76] zur Schallvisualisierung, Ortung und Quellentrennung mit Mikrofonarrays um. Jedes Einzelmikrofon nimmt Geräusche aus der Umgebung auf. Die analogen Signale werden in eine digitale Repräsentation konvertiert und gespeichert. Aufgrund der Laufzeitdifferenzen zwischen der Schallquelle und den einzelnen Mikrofonen lässt sich die ursprüngliche Position einer Schallquelle ermitteln. Wie in Abbildung 1 schematisch dargestellt, werden dabei für verschiedene Fixpunkte im Raum die einzelnen Aufnahmen zeitlich verzögert und addiert. Für einen Punkt (x;y) ist der zeitlich versetzte Signalverlauf immer gleich.

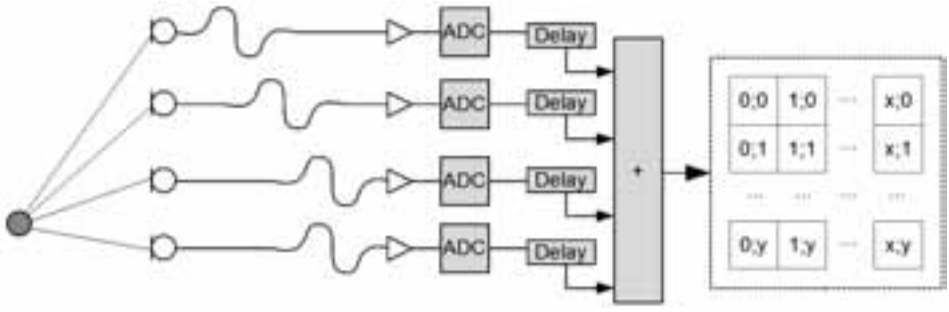

Abbildung 1: Hardware-Umsetzung des Algorithmus (v.l.n.r.): Schallquelle (rot), Mikrofonarray, verzögertes Signal, Vorverstärker, ACDs, Datenpuffer, Beamformer, Schallquellenvisualisierung

Die Addition der einzelnen Messpunkte ergibt in diesem Fall einen besonders starken Signalausschlag. Je unterschiedlicher die einzelnen Signale sind, umso geringer ist der

Signalausschlag nach der Addition. Im besten Fall löschen sich die Signale durch die Addition gegenseitig aus. Jedes Mikrofonsignal wird zunächst proportional zu einem maximalen Signalpegel über eine Vorverstärkerstufe skaliert und anschließend mit Hilfe eines ADCs in eine digitale Repräsentation konvertiert.

Grundlage für die Hardware-Umsetzung des Algorithmus ist ein FPGA-basiertes Spartan3E-Starter-Kit (Abbildung 2). Dieses wurde durch ein eigenes Hardware-Modul für die separate Verwendung von bis zu 24 Mikrofonen erweitert.

Abbildung 2: Spartan3E-FPGA und Mikrofonarray mit vier Einzelmikrofonen

Die auditive Lösung erlaubt eine nicht-invasive Ortung des aktuellen Sprechers. Aufgrund des sphärischen Erfassungsbereichs der Mikrofone ist neben der Lokalisierung von Dozierenden auch die Ortung von Sprechern im Publikum möglich.

2.2 Optische Ortung

Insbesondere in größeren Lehrveranstaltungen ist mit einem höheren Geräuschpegel als in kleineren Seminaren zu rechnen. Da die Beamforming-Methode mit zunehmendem Geräuschpegel zu ungenaueren Ergebnissen führt, wurde zusätzlich eine optische Ortung implementiert.

Zentrale Herausforderung bei der optischen Ortung von Personen ist die Identifizierung von menschlichen Körpern oder Körperteilen innerhalb eines größeren Bildes. Weit verbreitet ist die Identifizierung von Gesichtern als charakteristisches Merkmal. Selbst preisgünstige Digitalkameras sind heute in der Lage Gesichter zu identifizieren. Allerdings können diese Daten nicht von handelsüblichen Digitalkameras ausgelesen werden. Somit wären komplexe Eingriffe in die Kamera-Hardware erforderlich, die die Portierbarkeit der Lösung auf andere Hardware stark eingeschränkt hätte.

Eine weitere Methode verwendet Helligkeits- und Ähnlichkeits-Klassifikatoren (*Haar Classifier*) für eine deutlich schnellere und zuverlässigere Gesichtserkennung [LM02]. Durch die Trainierbarkeit derartiger Klassifikatoren sind zudem beliebige Objekte identifizierbar.

Auch im vorliegenden Projekt wurde die optische Gesichtserkennung über Haar Classifier realisiert. Durch OpenCV, eine freie Programmbibliothek mit umfassenden Algorithmen für die Bildverarbeitung, konnten Details des Algorithmus verborgen werden [BK08]. Die Portierung von OpenCV wurde bereits in vorangehenden Arbeiten beschrieben [Mue12]. Ein Training der Klassifikatoren für menschliche Gesichter war dank mitgelieferter Trainingsdaten nicht erforderlich.

Als Eingabegerät für die Bilddaten wurde eine handelsübliche Webcam verwendet. Die zur Aufzeichnung der Lehrveranstaltung eingesetzte Kamera kann somit nach Bedarf ausgetauscht werden, ohne die Funktionsweise der Kamerasteuerung zu beeinträchtigen. Die OpenCV-basierte Gesichtserkennung liefert 2D-Koordinaten der erkannten Gesichter innerhalb des von der Webcam aufgenommen Bildes. Diese werden über einen USB-Datenstrom angeboten, der zur weiteren Verwendung wie in Abschnitt 3 beschrieben in einen seriellen Datenstrom konvertiert wird.

Die optische Ortung hat im Gegensatz zur auditiven Ortung nur den kegelförmigen Erfassungsbereich der Webcam und ist somit nur zur Lokalisierung in einem vordefinierten Bereich geeignet (z. B. Bewegungsbereich des Vortragenden). Sie bietet aber dafür eine geringere Anfälligkeit gegenüber Störungen und kann zudem mit günstiger Standardhardware realisiert werden.

3 Steuerung der Kamera

Für eine zuverlässige Ausrichtung der Aufzeichnungskamera wurden beide Verfahren in Form eines Prototyps durch Studierende der Universität Potsdam im Wintersemester 2011/12 implementiert. Die Studierenden entwickelten in vier Gruppen vier verschiedene VHDL-basierte Software-Lösungen. Im Zentrum der gemeinsamen Architektur steht ein FPGA-Board. Dieses steuert zwei Servomotoren an, die die Kameraausrichtung horizontal und vertikal anpassen können.

Die Kameraausrichtung kann einerseits auf Grundlage der auditiven Ortung des Sprechers mittels Beamforming erfolgen. Alle angeschlossenen Mikrofone bilden zusammen ein Array als Grundlage für den Beamforming-Algorithmus. Ein zusätzlicher VGA-Ausgang wird für die direkte Visualisierung der lokalisierten Schallquellen verwendet.

Andererseits wurde die optische Ortung über OpenCV implementiert, allerdings im aktuellen Prototyp noch nicht auf das FPGA-Board portiert. Stattdessen leistet derzeit noch ein handelsüblicher PC die optische Erfassung durch eine Webcam und die OpenCV-basierte Auswertung der Bildsignale. Ein USB-Seriell-Konverter ermöglicht den Aufbau einer direkten, seriellen Kommunikation zwischen dem PC und dem FPGA. Dieser setzt die empfangenen Koordinaten in Steuersignale für die Servomotoren um. Die Portierung von OpenCV auf einen FPGA liegt bereits vor [Mue12], wurde aber bislang noch nicht in den Prototypen integriert. Damit entfällt der Host-PC in der Architektur des Gesamtsystems, und die Lösung wird portabel.

Abbildung 3: Der Prototyp zur automatischen Kamerasteuerung im Einsatz während einer Lehrveranstaltung.

Der in Abbildung 3 mit einem dreigliedrigen Mikrofonarray im Einsatz dargestellte Prototyp wurde in fünf realen Lehrveranstaltungen getestet. Diese fanden sowohl in größeren Hörsälen als auch in kleineren Seminarräumen statt.

Die Kamerasteuerung folgte dem Dozenten im Allgemeinen zuverlässig, wobei die auditive Ortung erwartungsgemäß in ruhigen Umgebungen deutlich besser funktionierte als in einer Umgebung mit hohem Hintergrund-Geräuschpegel. In ersten Tests störten auch die mechanischen Geräusche der Servomotoren die Ortung, woraufhin diese in größerer physischer Entfernung zum Mikrofon-Array montiert wurden.

Wie ebenfalls erwartet, erwies sich die optische Erkennung vor allem dann als zuverlässig, wenn der Dozent sich der Kamera zuwandte und den Erfassungsbereich nicht verließ.

Für die Weiterentwicklung der Lösung erscheint ein hybrider Ansatz aus auditiver und optischer Ortung sinnvoll. Dieser ist bereits im Rahmen einer Default- und Fallback-Ortung möglich, wird aber zu einer kombinierten Ortung mit gegenseitiger Verifikation weiterentwickelt.

4 Zusammenfassung und Ausblick

Bei der Aufzeichnung von Lehrveranstaltungen stellt der Beginn der Aufzeichnungskette – die eigentliche Aufnahme der Dozenten – eine bisher nur unzureichend betrachtete Herausforderung dar und birgt noch ein großes Automatisierungspotential. Dieser Artikel präsentiert ein System zur automatisierten Kamerasteuerung, das im Unterschied zu früheren Lösungen auf Basis eines Host-PCs unmittelbar zwischen ein handelsübliches Stativ und die Aufzeichnungskamera installiert wird und die Kamera nach erfolgreicher Ortung stets auf den Vortragenden zentriert. Zudem wurde ein Ansatz aus auditiver und optischer Lokalisierung prototypisch implementiert. Dabei sind im

Gegensatz zu verwandten Lösungen keinerlei Markierungen oder umfangreichen Vorkonfigurationen auf Seiten des Vortragenden erforderlich, so dass dessen Ortung völlig transparent erfolgt.

Das entwickelte System bietet dennoch einige Ansätze zur Weiterentwicklung. Beispielsweise wird die optische Lokalisierung bisher nur zweidimensional vorgenommen. Die zusätzliche Auswertung der Tiefeninformation würde auch ein Zoomen auf den Dozierenden ermöglichen und die Aufzeichnungsqualität weiter steigern. Allerdings stellt dies höhere Anforderungen an die zur Aufzeichnung genutzte Kamera. Dies widerspricht dem bisher verfolgten Ansatz der größtmöglichen Transparenz und Adaptivität. Zooming-Funktionalität zählt daher noch zu den offenen technischen Herausforderungen.

Weitere technische Herausforderungen betreffen die Portierung der optischen Ortungsalgorithmen auf das ressourcenarme FPGA-Board, die Implementierung eines Nutzerinterfaces zur Anzeige von Statusinformationen und diverse bauliche Systemoptimierungen und -minimierungen.

Im Moment wird zudem daran gearbeitet, die vorgestellte Lösung mit einem System zur Erfassung von Zeigegesten [Luc12] zu verbinden, damit beide das gleiche Equipment nutzen können. In der Summe beider Lösungen werden realitätsnahe E-Lectures erwartet. Der Mehrwert für die Studierenden soll dann im regulären Studienbetrieb evaluiert werden.

Denkbare Erweiterungen des Systems sind zudem Mechanismen zur Auflockerung der Aufzeichnung durch gelegentliche Schwenks & Zooms im Publikum oder die Anbindung von Systemen zum Management von Wortmeldungen, Umfragen, etc.

Literaturverzeichnis

[BK08] G. Bradski, A. Kaehler: “Learning OpenCV Computer Vision with the OpenCV Library”, O'Reilly, 2008

[BK76] J. Billingsley, R. Kinns: “The acoustic telescope”, in Journal of Sound and Vibration 04/48, Elsevier, 1976, S. 485-510.

[La+08] F. Lampi, S. Kopf, M. Benz, W. Effelsberg: “A Virtual Camera Team for Lecture Recording”, in IEEE MultiMedia 03/15, IEEE, 2008, S. 58-61.

[LM02] R. Lienhart, J. Maydt: “An Extended Set of Haar-like Features for Rapid Object Detection” IEEE ICIP, IEEE, 2002, S. 900-903.

[Luc12] U. Lucke: „Authentic Online Classes beyond traditional lecture streaming“, Keynote, in Proc. 5th e-Learning Baltics Conference (eLBa 2012), Fraunhofer Verlag 2012, S. 3-10.

[Mue12] F. Mühlbauer: „Entwurf, Methoden und Werkzeuge für komplexe Bildverarbeitungssysteme auf Rekonfigurierbaren System-on-Chip-Architekturen“, Dissertation, Universität Potsdam, 2012.

[TLH09] S. Trahasch, S. Linckels, W. Hürst: „Vorlesungsaufzeichnungen – Anwendungen, Erfahrungen und Forschungsperspektiven“, in i-com 03/08, München : Oldenbourg Verlag, 2009, S. 62.

[We+12] K. Weiß, D. Sailer and M. Braun: „Automated Recording of Lectures”, in Proc. 5th e-Learning Baltics Conference (eLBa 2012), im Druck.