Paper Review

„A Comparative Evaluation of Template and Histogram Based 2D Tracking Algorithms“

B. Deutsch, Ch. Gräßl, F. Bajramovic, J. Denzler

aus: W. Kropatsch et al., Pattern Recognition, 27th DAGM Symposium, Springer, 2005, (pp. 269-276)

Matthias Hillebrand
mhillebr@techfak.uni-bielefeld.de

im Rahmen des Seminars
„Visuelle Überwachung“

Angewandte Informatik
Technische Fakultät
Universität Bielefeld

Abstract

Das zugrunde liegende Paper vergleicht fünf aktuelle Regionen- und Histogramm-basierte Trackingverfahren, die in den Jahren 1998-2004 vorgestellt wurden, am Beispiel des Trackings von Personen. Untersucht wird die Performanz der Verfahren in Puncto Erkennungsrate und Geschwindigkeit. Als Testdatenbasis dienen frei verfügbare Beispiel-Videos von fest installierten Überwachungskameras [6].

Relevanz in Bezug auf das Seminar

Echtzeit-Tracking von Objekten oder Personen in 2D-Videobildern ist ein wichtiger Aspekt für viele automatische Überwachungssysteme. Es seien hier die Verfolgung von Personen, die Erfassung von Bewegungsmustern, Gesichtserkennung oder die Gesten-/Handlungserkennung genannt.
Die Autoren versuchen in diesem Vergleichstest, die Vor- und Nachteile der zwei Klassen von Trackingverfahren und auch der fünf einzelnen getesteten Verfahren herauszustellen. Dafür werden die Erkennungsraten und die Geschwindigkeit der Verfahren gemessen. Der Vergleich kann die Entscheidung für oder gegen den Einsatz eines der getesteten Verfahren erleichtern.

Zusammenfassung des Inhalts

In dem zugrunde liegenden Paper wird die Performanz von insgesamt fünf Trackingverfahren, davon zwei Regionen-basierte und drei Histogramm-basierte, verglichen. Als Datenbasis für den Test dienten sieben Überwachungsvideos von verschiedenen fest installierten Kameras [6]. In den sieben Videos wurden die Bewegungen von insgesamt zwölf Personen verfolgt. Jeder Tracker wurde zweimal mit jeder Person getestet. Der erste Durchlauf ging von der Annahme aus, dass eine reine Translation vorliegt, dass die Person ihre Entfernung zur Kamera also nicht verändert. Im zweiten Durchlauf wurden dann zusätzlich zur Translation auch die Skalierung betrachtet.

Regionen-basierte Trackingverfahren:
- Hager & Belhumeur [1]
- Hyperplane [2]

Regionen-basierte Verfahren beruhen auf Template-Matching. Dabei werden während des Initialisierungsschritts Bildpunkte aus dem Bild extrahiert, die zu dem zu trackenden Objekt gehören. Die Intensitäts- oder Farbwerte dieser Referenzregion bilden das Referenztemplate. In jedem Zeitschritt, also für jedes Bild, versuchen die Algorithmen nun, das Referenztemplate in der Umgebung der im vorherigen Bild gefundenen Position der Person zu finden. Dabei werden von den Verfahren verschiedene Transformationen des Templates wie Translation, Skalierung, Rotation oder affinen Transformationen berücksichtigt.

Histogramm-basierte Trackingverfahren:
- Mean Shift [3]
- Trust Region [4]
- Condensation [5]

Histogramm-basierte Verfahren beschreiben ein zu trackendes Objekt durch ein Histogramm, beispielsweise sein Intensitäts- oder Farbhistogramm. Das während der Initialisierung berechnete Referenzhistogramm wird für das Tracking mit den Histogrammen einzelner Bildausschnitte verglichen, die die alte Position der Person und eine kleine Bewegung in alle möglichen Richtungen abdecken. Die Histogramme der einzelnen Ausschnitte werden mit einem geeigneten Abstandsmaß mit dem Referenzhistogramm verglichen. Das eigentliche Tracking stellt somit ein Optimierungsproblem dar.

Für alle fünf Trackingverfahren des Vergleichstests wurden Personen durch rechteckige Bildausschnitte repräsentiert. Die exakten Positionen der Personen in einem Bild (Ground-Truth-Daten) liegen als Vergleichsdaten vor [6]. Bewegungen der Personen wurden als Translationen und Skalierungen dieser rechteckigen Bildausschnitte beschrieben. Jedes Verfahren wurde mit den Vergleichsdaten des Einzelbildes initialisiert, auf dem eine Person nach ihrem Auftreten zum ersten Mal verdeckungsfrei zu sehen ist.

Durch den Vergleich der Trackingergebnisse und der Vergleichsdaten wurde für jedes Bild der Distanzfehler e_c und der Regionenfehler e_r ermittelt. e_c stellt den euklidischen Abstand zwischen den Mittelpunkten der Rechtecke dar, während e_r definiert ist als der nicht-überlappenden Anteil der beiden Rechtecke:

$$e_r(A, B) := \frac{|A \setminus B| + |B \setminus A|}{|A| + |B|}$$

Dabei stellen A und B Mengen mit den Bildpunkten der durch die Ground-Truth-Daten beschriebenen bzw. der durch den Tracker ermittelten Bildregion dar.

Für die Messung der Geschwindigkeit der Tracker wurden für jedes Verfahren mit und ohne Skalierung die benötigte Zeit für die Initialisierung und die durchschnittliche Zeit pro Bild gemessen.

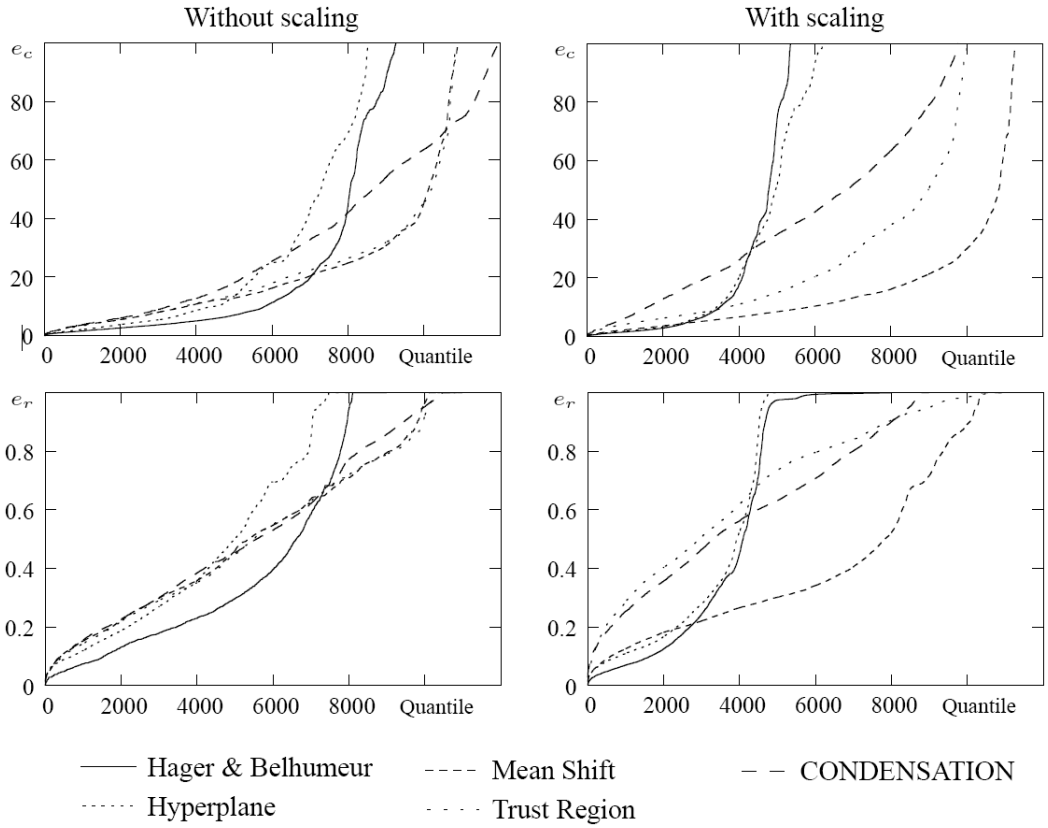

Abbildung 1: Ergebnisse des Trackervergleichs. Obere Reihe: nach Größe geordnete Distanzfehler e_c , untere Reihe: nach Größe geordnete Regionenfehler e_r . Linke Spalte: Translation ohne Skalierung, rechte Spalte: Translation mit Skalierung. (aus [1], siehe auch dort)

Die Auswertung der Messergebnisse wurde für jeden der beiden Tests (mit/ohne Skalierung) getrennt durchgeführt. Die Autoren haben für jedes der beiden Fehlermaße und für jeden Tracker die Fehlerraten der insgesamt ca. 12.000 Bilder aller zwölf verfolgten Personen zusammengefasst, aufsteigend sortiert und grafisch präsentiert (s. Abbildung 1).

Aus dieser Darstellung lassen sich einige Schlüsse ableiten: Zuallererst kann man sagen, dass die

Performanz stark zwischen den einzelnen Verfahren variiert, und zwar abhängig davon, ob Skalierung benutzt wird oder nicht und welche Fehlerrate betrachtet wird. Insgesamt gibt es keinen Tracker, der eindeutig der beste oder der schlechteste ist. Es lässt sich jedoch ein klarer Unterschied zwischen den Regionen- und den Histogramm-basierten Verfahren feststellen: Erstere haben in 40-60% (ohne Skalierung) bzw. 20% (mit Skalierung) der Bilder eine geringere Fehlerrate als die letzteren, danach steigt der gemessene Fehler jedoch sehr viel stärker an. Die Regionen-basierten erreichen die Fehler $e_r = 1$ und $e_c = 1$ (entspricht dem Verlust der Person) dadurch immer eher als die Histogramm-basierten. Abschließen kann man also sagen, dass die Regionen-basierten Verfahren für einen Teil der Bilder bessere Ergebnisse liefern, die Personen aber auch schneller verlieren, weshalb sie nicht so robust sind wie die Histogramm-basierten.

Diskussion

Das Paper liefert einen Vergleich einiger aktueller Tracking-Verfahren aus den Jahren 1998-2004. Gerade diese Aktualität macht es interessant für Wissenschaftler, die auf dem Gebiet des Personentrackings arbeiten.
Als kleine Schwäche empfand ich jedoch die relativ kleine Testdatenbasis von nur zwölf Personen in sieben Videos. Verliert ein Tracker auch nur eine einzige von den zwölf Person bereits zu Beginn und findet sie auch nicht wieder, so wird er dadurch im Vergleich zu den anderen im Ergebnis sehr massiv abgewertet. Zumindest heute (2006) wäre die zur Verfügung stehende Testdatenbasis [6] wesentlich größer als die vor zwei Jahren benutzen sieben Videos.
Ich nehme an, dass alle getesteten Verfahren ein Problem mit der Situation haben, wenn eine Person sich nicht zielstrebig und gerade durch das Bild bewegt, sondern ziellos hin- und herwandert und dadurch von verschiedenen Seiten zu sehen ist oder die Person mal von vorne und mal von oben zu sehen ist. Speziell die Regionen-basierten Tracker dürften dann Schgwierigkeiten bekommen, ihr Referenztemplate wiederzufinden.
Das Paper ist meiner Meinung nach völlig eigenständig, da nicht die Entwicklung oder Weiterentwicklung eines bestehenden Verfahrens Gegenstand ist, sondern der Vergleich von fünf existierenden, aktuellen Verfahren. Auch die Evaluationsmethode scheint von den Autoren selbst entwickelt worden zu sein, da sie sich nicht auf eine andere Veröffentlichung berufen. Genau diese Evaluation hat aber auch ihre Schwächen. Da die Bilder aller zwölf Personen in einen Topf geworfen und nach ihrem Erkennungsfehler sortiert werden, erfährt man nichts darüber, wie der Tracker sich auf einer einzelnen Bilderfolge verhält und wie lange es durchschnittlich dauert, bis er eine Person verliert.

Bei der Performanz in puncto Geschwindigkeits lässt sich kein eindeutiges Urteil fällen: Einzig und allein das Hyperplane-Verfahren, das in in ~0.5 sec initialisiert und das CONDENSATION-Verfahren, das in ~0.1 sec trackt, fallen aus dem Rahmen. Alle anderen Zeiten bewegen sich im Bereich von ~1-10 ms, sowohl für die Initalisierung als auch für die Tracking-Zeit für ein Bild.

Fazit

Anhand der Testergebnisse sieht man, dass das Problem, eine Person in einer realen Umgebung mit inhomogenem Hintergrund zu tracken, nicht trivial ist. Zumindest unter den fünf vorgestellten Tracking-Verfahren (und es ist davon auszugehen, dass das auch für alle hier nicht getesteten gilt) gibt es nicht das eine Verfahren, das die Antwort auf alle Fragen und Probleme ist.

Literatur

1. Hager, G., Belhumeur, P.:
   Efficient region tracking with parametric models of geometry and illumination.
   IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1998) 1025–1039

2. Jurie, F., Dhome, M.:
   Hyperplane approach for template matching.
   IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (2002) 996–1000

3. Comaniciu, D., Meer, P., Ramesh, V.:
   Kernel-Based Object Tracking.
   IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (2004) 564–575

4. Liu, T.L., Chen, H.T.:
   Real-Time Tracking Using Trust-Region Methods.
   IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (2004) 397–402

5. Perez, P., Hue, C., Vermaak, J., Gangnet, M.:
   Color-Based Probabilistic Tracking.
   In: 7th European Conference on Computer Vision. Volume 1. (2002) 661–675

6. CAVIAR: EU funded project, IST 2001 37540,
   URL: http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ (2004)