DISS. ETH Nr.

Strategies of Robust Object Recognition for the Automatic Identification of Human Faces

A B H A N D L U N G

Zur Erlangung des Titels

DOKTOR DER NATURWISSENSCHAFTEN

der

EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE ZÜRICH

vorgelegt von

MARTIN BICHSEL
Dipl. Phys. ETH
geboren am 21. Mai 1962
von Sumiswald, BE

Angenommen auf Antrag von:
Prof. O. Kübler, Referent
Dr. P. Seitz, Korreferent

O. Kübler

Zürich, 1991

Deutsche Zusammenfassung

Wie leicht und zuverlässig wir menschliche Gesichter erkennen können, ist eine der erstaunlichsten Fähigkeiten des menschlichen Sehsystems. Um menschliche Gesichter oder andere natürliche Objekte erkennen zu können, wird eine besonders robuste Art der Informationsverarbeitung benötigt, die nach unserem Wissen bis jetzt noch von keinem Signal- oder Bild-verarbeitungsalgorithmus gezeigt wurde.

Diese Arbeit hat gezeigt, dass automatische Gesichtserkennung und allgemeine Objekterkennung stark von einer interdisziplinären Problembehandlung profitieren können. Gesammeltes Wissen aus Neurobiologie und Psychologie über den Aufbau biologischer visueller Systeme wurde mit Resultaten aus Signalverarbeitung, Informationstheorie und Gruppentheorie kombiniert und erlaubte den Aufbau eines kompletten Gesichtserkennungs-Programmes. Ohne jegliche menschliche Intervention werden Gesichter von einer Video-Kamera aufgenommen, lokalisiert und identifiziert.

Die grundlegenden Aufgaben beim Erkennen von Objekten und Gesichtern, z.B. ein Gesicht in einer Bildszene zu lokalisieren, sind Klassifikations-Aufgaben. Daher bieten sich geschichtete neuronale Netzwerke als universelle Lösung an. Das Forschungsgebiet Neuroinformatik leidet jedoch an zwei wichtigen Schwachpunkten: (1) Es gibt kein universelles Mass, um die Leistungsfähigkeit eines ganzen Netzes oder seine Teile zu burteilen und (2) alle bekannten Lernalgorithmen erreichen nur ein sub-optimales Training des Netzes. Die neue Sicht von neuronalen Netzwerken, die im Laufe dieser Arbeit entwickelt wurde und die neuronale Netzwerke als Mehr-Stufen-Kodierer (im Sinne von Shannons Informationstheorie) betrachtet, hat zu Lösungen für beide Probleme geführt: Die bedingte Klassen-Entropie wird zum Mass der Wahl um die Leistungsfähigkeit eines Netzwerkes zu beurteilen und ein neuer Lern-Algorithmus, der diese Entropie mit Hilfe von "Simulated Annealing" minimiert, erlauben eine Generierung von optimalen Netzen für beliebige Klassifikations-Probleme.

Mit einer geometrischen Interpretation von neuronalen Netzwerken in mehrdimensionalen Räumen wird gezeigt, dass Algorithmen zur Signal-Vorverareitung existieren könnten, die zu einer wesentlichen Leistungssteierung der Netwerke führen, indem die Klassen-Grenzen im Parameter-Raum geglättet werden. Bei Objekterkennungs-Problemen weisen die Klassen-Grenzen typischerweise eine Feinstruktur auf, die von den verschiedenen möglichen Transformationen der Objekte herrührt. Mit Hilfe der Gruppentheorie können die relevanten Grössen - invariante Mengen - für Rotationen oder andere wichtige Transformationen bestimmt werden. Untermauert vom Wissen über das menschliche Sehsystem führt dies zu einer Vorverarbeitung der Input-Szene, die aus einer räumlichen Tiefpass-Filterung (Pyramide) besteht, von der die lokale Orientierungs-Information auf jedem Auflösungs-Niveau gewonnen wird.

Diese neue Bilddarstellung erhöht die Leistungsfähigkeit von neuronalen Netzen und herkömmlichen "matched filters" wesentlich. Als Bsp. wurde ein Netz trainiert, menschliche Gesichter in vorverarbeiteten Bildern zu detektieren. Das resultierende Netzwerk detektierte Gesichter in einer komplexen Szene, die keines der gelernten Gesichter enthielt, mit einer hohen Treffsicherheit: Im Bild einer Menschenmenge wurden weniger als 0.06% der Bild-Pixel fälschlich als Gesicht klassifiziert und mehr

als 80% der Gesichter wurden korrekt lokalisiert.

Komplementär zu den neuronalen Netzwerken wurde ein neuer Objekterkennungs-Algorithmus entwickelt, der eine Sequenz von Fokus-Punkten durchläuft in Analogie zu den sakkadischen Augenbewegungen des menschlichen Sehsystems. In jedem Fokus-Punkt werden die lokal invarianten Bildmerkmale gewonnen. Diese werden benutzt, um die Schätzwerte der Objektparameter (Position und Orientierung) zu verbessern aund inkonsistente Objekt-Interpretationen mit einem minimalen Rechenaufwand zu elliminieren. Mit diesen Algorithmus wird ein bestimmtes Werkzeug in einem Haufen anderer Werkzeuge oder ein menschlicher Kopf mit unbekannter Orientierung und Position zuverlässig lokalisiert, indem nur 11 lokale Fokus-Punkte angeschaut werden.

Basierend auf Resultaten aus der Psychologie werden zwei komplementäre Strategien für die Identifikation menschlicher Gesichter untersucht: "Matched filters" (einschichtige neuronale Netzwerke) werden verwendet um gute Übereinstimmungen zwischen den wichtigsten Gesichtsteilen eines untersuchten Gesichtes und den entsprechenden Teilen von Personen in einer Datenbasis zu finden. Es wird die grobe Kopfform, die Augenregion und die Mund/Nasen-Region verglichen.

Die Resultate dieses "matched filtering" werden durch eine Technik verfeinert, die ursprünglich vom französischen Kriminologen Bertillon entwickelt wurde. Wichtige Gesichtspunkte, z.B. Augenecken, Nasenspitze und Mundwinkel, werden mit "matched filtering" vermessen und geometrische Beziehungen zwischen den Punkten berechnet. Der resultierende mehrdimensionale Vektor, der die geometrischen Beziehungen zwischen den Gesichtwpunkten enthält, beschreibt die Geometrie eines Gesichtes. Dieser Vektor wird mit der intra-personalen Varianz der verschiedenen Gesichtsteile normiert. Dieses Vorgehen verringert die Empfindlichkeit der Klassifizierung für Gesichtsteile die bei einer Person stark variieren, da sie z.B. stark von der Kopforientierung abhängen. Der normierte Vektor wird dann dazu benutzt um anhand der minimalen euklidischen Distanz die endgültige Zuweisung des vorliegenden Gesichtes zu einer Person in der Datenbasis vorzunehmen.

Mit den beschriebenen Komponenten für die Lokalisierung und Identifikation eines menschlichen Gesichtes wurde ein kompletes Gesichtserkennungs-System realisiert um die Leistungsfähigkeit in einer praktischen Anwendung zu testen: der elektronische Pförtner. Dieses Demonstrations-Experiment wurde als automatische Zutrittskontrolle, die rein auf der automatischen Gesichtserkennung basiert, entwikkelt und aufgebaut. In einem Testset, bestehend aus 397 Gesichtern von 70 verschiedenen Personen, wurden 90% der bekannten Gesichter korrekt erkannt und 50% der fremden Gesichter falsch akzeptiert, für einen bestimmten Kompromiss zwischen hoher Erkennungsrate und niedriger Falsch-Akzeptierrate. Im Gegensatz zu anderen Studien wurden diese Personen nicht speziell ausgewählt, so dass verschiedene praktische Schwierigkeiten, wie z.B. Brillen, Bärte, wechselnde Frisuren, etc., zu bewältigen waren.

Mit dieser Demonstration haben wird die Realisierbarkeit der Gesichtserkennung in einer realistischen Umgebung gezeigt. Viele dieser neu entwickelten Algorithmen sind auch auf allgemeine Objekterkennungs-Probleme anwendbar. Dies könnte eine Reihe von neuen praktischen Anwendungen in der Industrie ermöglichen.

English Summary

The ease and the reliability with which people can recognize and identify human faces is one of the most amazing capabilities of the human visual system. In order to recognize human faces or other natural objects, a particularly robust information processing scheme is required, to our knowledge as yet not demonstrated with any signal and image processing algorithm.

This work has shown that automatic face recogniton, and object recognition in general, can profit a lot from an interdisciplinary approach. Knowledge gained in neurobiology and psychology about the elements of biological vision systems was employed, together with results in signal processing, information theory and group theory and lead to the development of a complete face recognition program. Without any operator intervention it localizes and identifies human faces in images captured with a video camera.

Since the basic problems in object and face recognition, e.g. the localization of human faces in a scene, are difficult classifying problems, feedforward neural networks could lend themselves as the universal solution for these applications. Today's research field of neurocomputing, however, suffers from two main deficiencies: (1) there is no universal measure for judging the performance of a neural network or parts of it, and (2) all known teaching algorithms achieve only sub-optimum training of the networks. The novel view of neural networks – developed in the course of this work – as multi-stage encoders in the framework of Shannon's information theory, leads to solutions of both these problems: The conditional class entropy as the measure of choice for judging the performance of any portion of a neural network, and a novel teaching algorithm, minimizing the conditional class entropy using simulated thermal annealing, which is capable of generating optimum neural networks for any type of classification problem.

Employing a geometric interpretation of neural networks in multidimensional spaces it is shown that signal preprocessing algorithms could exist, capable of substantially increasing the performance of neural networks by efficiently warping the signal space. In object recognition the signal space is typically structured on a fine scale due to various possible transformations of an object. Group-theory allows to reveal and classify the relevant quantities - invariant sets - for important transformations, such as rotation. Underpined by biologists' knowledged about the human visual system this leads to an image preprocessing consisting of the computation of a spatially low-pass filtered multi-resolution representation (pyramid) of the input scene, from which the information about local orientation is derived at each level of resolution.

This new image representation substantially increases the performance of neural networks *and* traditional matched filters. As an example, a neural network was trained to detect human faces in preprocessed images. The resulting neural network showed good face-discrimination properties when presented with complex scenes containing other human faces than the ones with which it was trained: In a busy scene – for example the image of a crowd – less than 0.06% of the pixels were incorrectly classified as faces (false positive), and more than 80% of the faces were correctly localized.

Complementary to the neural networks approach an efficient new object recognition algorithm was developed which steps through a sequence of focal points, where local invariant image characteristics are extracted, similar to the saccadic eye movements of the human visual system. The local information at each focal point is exploited in order to update the object parameters - its position and orientation - and to eliminate inconsistent object interpretations with a minimum amount of computations. As examples a particular tool in a pile of other tools or human heads with unknown orientation and position are reliably localized by looking at only 11 local focal points or less.

Based on the results of cognitive psychologists, two complementing strategies for the identification of human faces are investigated: Matched filters (single layer neural network classifiers) are used – again on the preprocessed representation of the scene – to find good matches between the most important features of a face under investigation and the corresponding features of persons in a database. The overall form of the head, the eye region, and the nose/mouth region are compared.

The matching results are refined by employing a technique pioneered by the French criminologist Bertillon, in which landmark features in a face are localized – again by using matched filters – and geometric relationships between them are calculated. These features include corner points of the eyes, the tip of the nose, the corner points of the mouth, etc. A multidimensional vector of the geometric relationships of the feature points results which describes the geometry of a face. This vector is then normalized with the intra-person variances of the different features, in order to reduce the sensitivity of the final classification process to features that vary substantially in the same person (e.g. depend heavily on the orientation of the head). The normalized vector is then used with a traditional minimum Euclidean distance classifier to perform the final identification of the presented face as a particular person.

With the described components for localizing and identifying a human face a complete system has been set up to test its performance in a practical realization: the electronic concierge. We have designed, constructed and set up this demonstration experiment, which could serve as an automatic access control system to our laboratory, based solely on face recognition. The face images with which the recognition rate were then determined were not part of the training set, hence reflecting the practical conditions under which a realistic demonstration must take place. Based on a test set of 397 faces belonging to 70 different people we observed a correct identification rate of 90% and a false access rate of 50%, selecting a particular tradeoff beween correct identification rate and false access rate, for a particular tradeoff between a high identification rate and a low false access rate. In contrast to other studies these 70 persons were not specially selected, thus introducing such practical difficulties as glasses, beards, changing hairstile, etc.

With this demonstration we have shown the feasibility of face recognition in real-world environments using the techniques and methods developed during the course of this work. Several of these newly developed algorithms developed are also suited for general object recognition problems. This could open up many new applications in industry and help solve difficult practical problems.