PARAVISION

March 2023

Paravision Search

Scalable, Elastic, Enterprise-Grade Face Recognition

Paravision Search takes Paravision Face Recognition, which has been available until now as an SDK or discrete docker-based matcher, and integrates it into a true enterprise-grade search system. This includes massive scalability, supporting both very large databases and extreme request concurrency. It also includes native elasticity, allowing Paravision partners to adjust compute resources to meet real-time demands for peak traffic or quiet periods.

Paravision Search integrates sophisticated gallery management and dynamic, custom attribute filtering, enabling definition of sub-galleries and focused search. This allows for better speed and accuracy optimization when full database search isn't necessary: for instance, when date, geography, or authorization can be added as search criteria for common customer experience use cases like air travel, stadium access, or payments. So, while Paravision Search scales easily to handle hundreds of millions of records, it is equally optimized for high-concurrency, smaller database applications like day-of-travel or day-of-event.

Meanwhile, Paravision Search is designed to be easy-to-deploy and manage for systems administrators. Integrated health management, performance monitoring, and other dashboards deliver the modern user experience IT professionals demand.

Supported Features

- Biometric template and image matching
- Rapidly configurable, highly flexible architecture supporting wide range of use cases
- Dynamic sub-gallery creation and search
- Extensible attribute-based search filters
- Programmable, API-controlled elasticity
- REST and gRPC interfaces
- Runs on Kubernetes and is cloud-ready

Performance Metrics

- Gallery size scales to hundreds of millions of records while still fully optimized for smaller galleries
- Enrolls 50 million identities in less than 3.5 hours
- 1.4 million matches per second per core on Intel(R) Xeon(R) @ 3.10GHz
- 120 concurrent requests per second with 16 CPUs and 100,000 templates
- Optimized for Paravision Gen 5 face recognition, ranked #1 globally in NIST FRVT 1:N (Visa-Border, Feb 2022)

Paravision Face Recognition

Paravision Search is powered by Paravision Face Recognition, which has been repeatedly recognized by NIST FRVT as a top global provider and the most accurate U.S.-based face recognition technology provider across all use cases. This includes 1:1 verification, 1:N identification with multi-million record databases, matching against hundreds to tens of thousands for airport day-of-travel, matching against a wide range of demographic groups, as well as handling the full range of image quality characteristics that range from passport-quality to fully unconstrained quality images.

Commitment to Open Systems Architecture

Paravision Search doesn't replace Paravision's other face recognition products. It is simply a more sophisticated embodiment of the same technology for partners looking for a fully-integrated, enterprise-grade face search capability out-of-the-box. Paravision remains fully committed to an Open Systems Architecture, and to offering partners SDKs, discrete Docker containers, or full Search capabilities, all of which are API-driven. Paravision powers its partners with world-class face recognition technology in a way that fits their technical expertise and deployment goals, supporting transformative solutions for the next generation identity, security, efficiency, and user experiences.

Beyond Faces

While Paravision Search is focused today on enabling enterprise-grade face matching, the foundational enabling technology is AI vector search. With this in mind, Paravision Search can be rapidly adapted to other image search applications. Contact us for more information about image search beyond faces.

Supported Cloud Platforms and Computing Environments

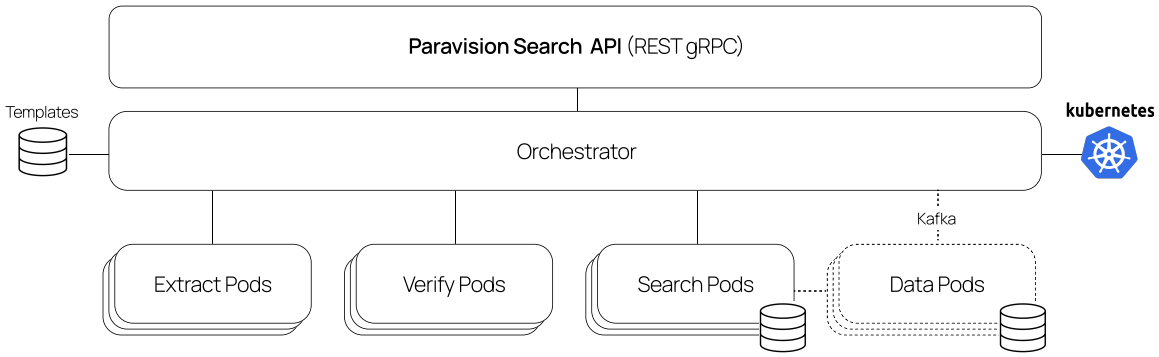
Paravision Search utilizes the latest technologies, allowing for flexibility in deployment methodologies. Search can be deployed on-premise or within a cloud service. Moreover, Paravision Search cloud configuration enables you to use the efficiency of elasticity from a cost and performance perspective.

On-Premises / Private Cloud	Google Cloud Platform	Amazon Web Services	Microsoft Azure
			

Technical Specifications

Deployment method	Docker containers using Kubernetes, supporting on-premises or cloud-based deployment (including GCP, AWS, Azure)
Supported operating systems	Windows: WSL and Docker Desktop Linux: Ubuntu 22.04
Supported template generation platform	Intel CPU (OpenVINO) NVIDIA GPU (TensorRT)
Client interfaces	REST API gRPC
Supported probe types	Image Template
Supported functions	1:N identification 1:1 verification Few:few verification Template generation Template + metadata export
Performance	Scalable to meet any DB size, concurrency, and latency requirements • Minimum latency: 100 msec • DB size: Unlimited • Concurrency: Unlimited
Elasticity	Automated / API-programmable / dynamically adjustable to meet system performance requirements
Search Optimization Features	Dynamic sub-gallery creation Filtering with customizable, extensible, dynamic attributes, including multiple attributes
Monitoring	Grafana, Loki, Prometheus
Database choices	Postgres (recommended) MariaDB MySQL Oracle MSSQL
Internal data synchronization	Data-sync via gRPC (default) Data-sync via Kafka (optional)

System Architecture



Performance

Paravision Search allows customers to input desired SLA and concurrency requirements, and the system will intelligently determine the number of nodes required to meet the expected performance. In a production deployment, the type of hardware will influence the number of nodes necessary for a use case. The below table captures the 95th percentile of matches per second to match a template on different types of GCP (Google Cloud Platform) and AWS (Amazon Web Servers) instances. It demonstrates the matching speed for Paravision's precision template, helping guide partners in picking the appropriate hardware for deployment.

Instance Type	vCPU	CPU Family	P95 Matches/ Sec/Core
GCP: c2-standard-8	8	Intel Xeon CPU @ 3.10GHz	1.4M
GCP: c2-standard-16	16	Intel Xeon CPU @ 3.10GHz	1.3M
GCP: c2-standard-60	60	Intel Xeon CPU @ 3.10GHz	411K
AWS: c4.4xlarge	16	Intel® Xeon® CPU E5-2666 v3 @ 2.90GHz	777K
AWS: c6in.4xlarge	16	Intel® Xeon® Platinum 8375C CPU @2.90GHz	1.4M
AWS: c6in.8xlarge	32	Intel® Xeon® Platinum 8375C CPU @ 2.90GHz	1.2M
AWS: c5.metal	96	Intel® Xeon® Platinum 8275CL CPU @ 3.00GHz	479K
AWS: c6a.48xlarge	128	AMD EPYC 7R13 Processor	67K