Atom 프로세서 기반의 모바일 플랫폼을 위한 인식 하드웨어 가속기

이승은 (서울과학기술대학교)

Ⅰ. 서 론

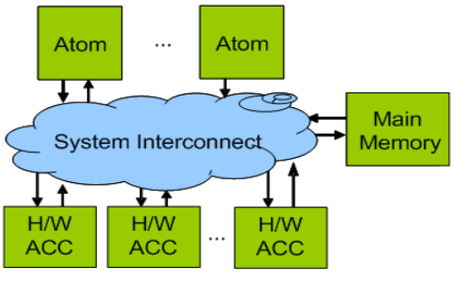
스마트 폰 및 모바일 인터넷 기기 (Mobile internet devices)의 대중화와 함께, 사용자의 편의를 위한 입출력 기기 기술이 개발되고 있다. 이러한 지능형 입출력 기기는 이미지, 음성, 제스처 등의 인식을 기반으로 하고 있고, 빠른 시일 안에 모바일 디바이스에 적용될 것으로 예상되며, 제품들이 출시되고 있다. 예를 들면, Mobile Augmented Reality (MAR)는 사용자가 관심을 가지고 있는 물체의 영상을 모바일 단말기의 카메라로 획득하면, 모바일 단말기가 이를 인식하여 관련된 정보를 사용자에게 실시간으로 제공하는 응용제품이다[1]. 또한, 음성인식을 이용한 받아쓰기, 인터넷 검색, 및 메시징 서비스가 기존의 키패드를 이용한 타이핑 방법을 대체하고 있다. 이는 모바일 단말기의 영상 및 음성 인식의 실시간 처리를 위한 고속 연산 및 저 전력 소모를 요구한다. 영상 및 음성 인식을 위해서는, 단말기가 대상 데이터(이미지, 음성 또는 문자 등)로부터 가장 관련 있고 적절한 특징 정보를 추출해 내야하며, 이러한 특징 정보를 기반으로 데이터베이스에 가지고 있는 정보와 비교하는 등의 인식 기능을 지원해야 한다. 정확한 인식을 위해서는 많은 연산을 요구하며, 이를 모바일 단말기에서 처리하기 위해서는 응용제품의 워크로드를 분석하고 이해하여 플랫폼의 구조를 적절하게 결정해야 한다.

내장된 프로세서 각각의 장점들을 전체 시스템의 성능을 향상시키는데 사용할 수 있기 때문에, 최근 Multi Processor System-on-Chip (MPSoC)이 많은 연산을 요구하는 응용 제품군에 적용되고 있다. 일반적으로, 많은 연산을 요구하는 기능들은 하드웨어 가속기에 할당되고, 그 외의 작업 및 제어는 내장된 범용 CPU가 수행한다.

본고에서는 인텔사의 저 전력 Atom 프로세서를 내장한 모바일 플랫폼에서의 SURF/ OpenCV[2] 기반의 영상인식과 Sphinx[3]를 기반으로 한 음성인식의 워크로드를 분석하고, 고속 연산을 요구하는 기능 블록을 하드웨어 가속기로 설계하여, MPSoC의 구조를 제안하고 그 성능을 보인 모바일 플랫폼을 위한 하드웨어 가속기 설계 사례를 소개 한다[4,5].

Ⅱ. Multi-processor System-on-Chip

MPSoC는 각각 다른 기능을 하는 다수의 프로세서를 집적하여 각각의 장점들을 이용하여 전체 시스템의 성능을 향상시킬 수 있어, 고속연산을 요구하는 응용제품에 많이 응용되고 있다. 특히 모바일 플랫폼에서의 내장 프로세서는 그 성능이 제한적이기 때문에 하드웨어 가속기를 이용한 성능 향상은 매우 효율적이다. 보통 고속 연산을 요구하는 작업은 하드웨어 가속기로 구현되고 시스템 제어 등을 위한 작업은 내장 프로세서에서 처리된다. 〈그림 1〉은 하드웨어 가속기를 내장한 모바일 플랫폼의 블록도이다. 아톰 프로세서의 소프트웨어는 하드웨어 가속기와 협업하여 고속 연산이 필요한 응용제품의 기능을 실현한다.



〈그림 1〉 MPSoC 구조

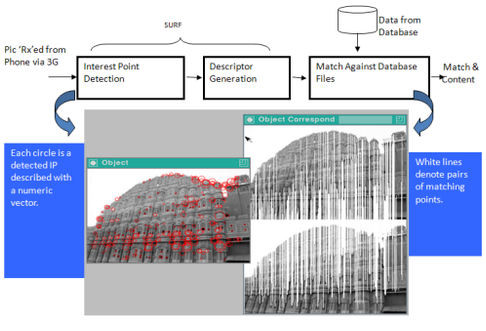
Ⅲ. Mobile Augmented Reality

영상인식은 다양한 응용 제품에 적용되고 있으며, 모바일 기기 및 게임 플랫폼에서의 새로운 입력 장치로 진화하고 있다. 인텔은 스마트 폰 및 모바일 인터넷 단말기에서의 Mobile Augmented Reality (MAR) 실시간 구현을 위하여, 워크로드를 분석하고 두 개의 하드웨어 가속기를 포함한 MPSoC 구조를 제안하였으며[4], 이를 다수의 소형 코어를 내장한 서버 구조로 확장시켰다[5].

〈그림 2〉는 사용자가 촬영한 입력 이미지로부터 영상인식을 완료하기까지의 MAR 영상인식 흐름을 보여준다. 이는 입력된 이미지와 데이터베이스의 이미지들과의 비교를 통해 인식을 완료하게 되며, 크게 다음 세 개의 기능 블록으로 구현된다.

- Interest-point detection: 입력된 이미지에서 interest-point를 찾아낸다.
- Descriptor generation: 추출된 interest-point로부터 descriptor 벡터를 생성한다.
- Match: 입력 이미지의 descriptor 벡터들과 데이터베이스의 descriptor벡터들을 비교하여 인식을 완료한다.

〈그림 3〉(a)는 인텔의 i7과 Atom 프로세서에서 소프트웨어로 구현된 MAR 연산시간을 나타낸다. 각각 프로세서에서의 소프트웨어 최적화 (Atom(Opt) 와 Nehalem(Opt))는 연산시간은 2배정도 단축시키며, i7와 Atom에서의 연산시간은 약 4배정도 차이가 있다. 〈그림 2〉(b)는 Atom 프로세서에서 각각 연산 블록의 신호처리 시간을 나타내며, Interest-point detection과 Match 프로세싱이 많은 연산 시간을 요구하는 hotspot 임을 보인다. 또한 데이터베이스의 이미지 수가 증가할수록 Match 프로세스의 연산 시간은 크게 증가한다. 이에 두 연산을 각각 처리하는 하드웨어 가속기를 설계한다.

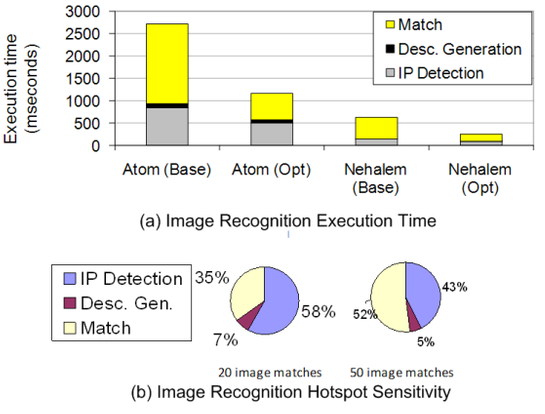


〈그림 2〉 Image Recognition Flow[4]

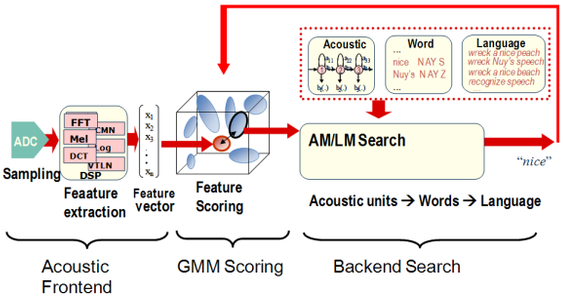
Ⅳ. Speech Recognition

음성인식은 다양한 휴대기기의 입력 수단으로 사용될 수 있으며, 사용자가 편하게 사용할 수 있는 기능이다. 이에, 많은 수의 어휘를 지원하며 연속 음성의 인식을 실현하고자 하는 연구가 지속되고 있으나, 이를 휴대 단말기에 구현하기에는 아직 어려움이 있다. Sphnix3[3]는 1) acoustic front end, 2) Gaussian mixture model (GMM) scoring, 그리고 3) acoustic/language model search의 3개의 연산으로 구현되는 음성인식 알고리즘이다(그림 4 참조).

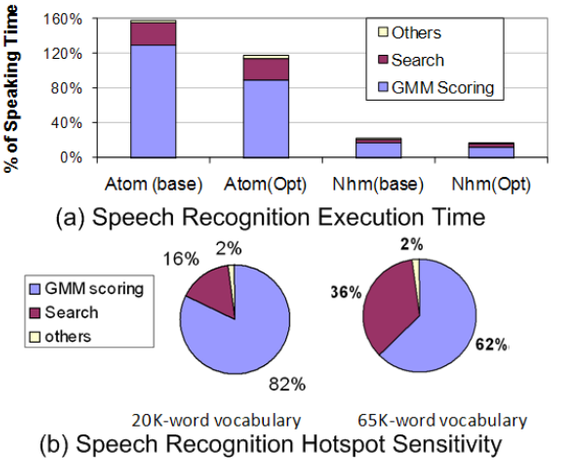
〈그림 5〉는 인텔 i7과 Atom 프로세서에서, Sphnix3를 소프트웨어로 구현하여 그 성능을 Wall street journal의 음원으로



〈그림 3〉 MAR execution time and primitive hotspots[5]



〈그림 4〉 Speech Recognition Flow



〈그림 5〉 Sphinx3 execution time and primitive hotspots[5]

분석한 결과이다. 약 70%의 시간이 GMM scoring에 소요되며, 잡음 환경에서는 그 연산 시간이 크게 늘어난다. 이를 해결하기 위하여, GMM scoring 하드웨어 가속기를 구현하여 Atom 프로세서를 기반으로 하는 플랫폼에서 음성인식을 구현한다.

Ⅴ. Recognition Server Architecture

Ⅲ장과 Ⅳ장에 기술된 영상 및 음성인식 알고리즘의 소프트웨어 구현 결과를 바탕으로, 모바일 플랫폼을 위한 실시간 인식 프로세서의 설계를 위하여 Atom 프로세서를 주 CPU로 하며, 3개의 하드웨어 가속기를 내장한 프로세서 구조를 제안하고 그 성능을 분석하였다. 제안된 플랫폼은 인식 응용제품을 위하여, 1) 충분한 인식률을 실시간으로 제공하며, 2) 전력 소모를 줄이고, 3) 응용제품 개발자가 프로그램을 편하게 할 수 있도록 하는데 중점을 두고 개발되었다.

1. Hardware Accelerators

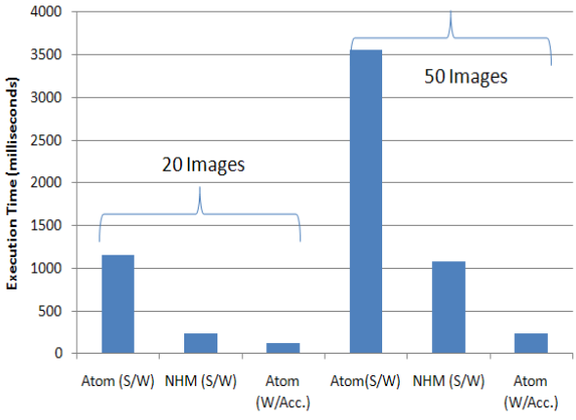
영상 및 음성인식을 저 전력 프로세서(인텔 Atom)에서 실시간 구현을 위하여 다음의 하드웨어 가속기를 설계하였다 (그림 6 참조).

- Match 가속기: 영상인식 하드웨어 가속기로 입력 이미지의 descriptor 벡터와 데이터베이스의 descriptor 벡터들과의 거리를 병렬로 계산하고, 가장 많이 매치된 이미지를 인식 결과로 출력한다.
- IPD 가속기: 입력 이미지로부터 Hessian 행렬을 계산하여 interest-point를 추출한다.
- GMM 가속기: 매 프레임의 음성 신호에 대해서 GMM scoring을 수행하는 하드웨어 가속기로 그 구조가 Match 가속기와 비슷하다.

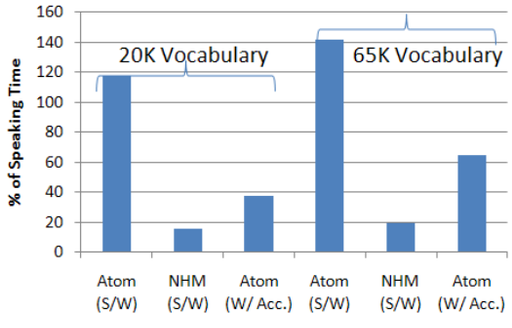
2. Performance Benefits

제안된 하드웨어 가속기는 Verilog HDL로 구현되었으며, FPGA 플랫폼에서 검증되었다. 또한, 45nm 공정으로 합성되어 그 성능, 전력 소모 및 회로크기를 산출하였다.

〈그림 7〉은 제안된 하드웨어 가속기를 포함한 영상 및 음성인식의 처리시간을 보인다. 하드웨어 가속기를 내장한 Atom 프로세서는 영상처리의 경우 Nehalem프로세서와 비교하여 약 2배 (20개의 DB 이미지) 및 약 5배 (50개의 DB 이미지)의 성능을 보였다. 음성인식의 경우, GMM scoring 가속기를 내장한 Atom 프로세서의 성능이 Nehalem 프로세서보다 2-3배 느렸으나, 그 전력소모는 1/200정도였으며, 실시간 처리를 하는데 충분하였다. 자세한 성능 비교 결과 및 분석은 [5]에 기술되어 있다.

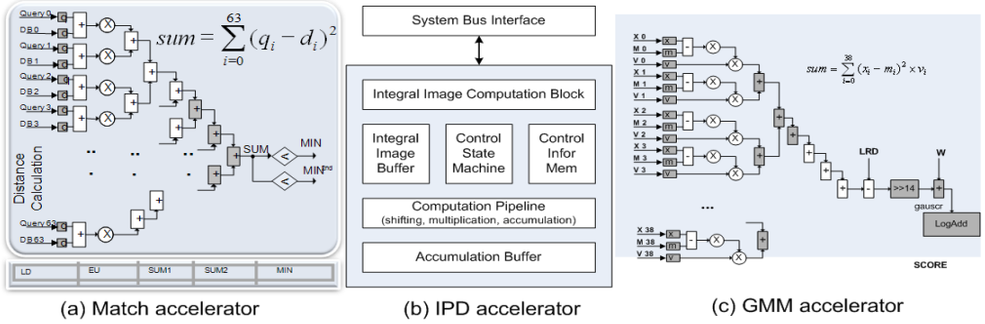


(a) 영상인식의 성능 개선



(b) 음성인식의 성능 개선

〈그림 7〉 3가지 플랫폼에서의 영상 및 음성인식 성능 개선[5]



(a) Match accelerator (b) IPD accelerator (c) GMM accelerator

〈그림 6〉 제안된 하드웨어 가속기 구조

Ⅵ. 결 론

본고에서는 모바일 플랫폼을 위해 개발된 하드웨어 가속기를 내장한 인텔 Atom 기반의 영상 및 음성인식 설계 사례를 살펴보았다. Small core인 Atom 프로세서를 고속의 연산 및 저 전력 소모를 요구하는 응용제품에 적용하기 위하여, 워크로드 분석을 수행하였으며, 많은 연산을 요구하는 블록을 하드웨어 가속기로 설계하여 그 성능을 분석하였다. 이러한 하드웨어 가속기를 내장한 프로세서가 향후 모바일 플랫폼에서 각광받을 것으로 예상하며, 추후 새로운 응용 제품에 대한 워크로드 분석 및 소프트웨어와 하드웨어 가속기 간의 연동을 편하게 할 수 있는 구조 연구가 계속 되어야 할 것이다.

참 고 문 헌

[1] G. Takacs et. al., “Outdoors Augmented Reality on Mobile Phone using Loxel-Based Visual Feature Organization,” ACM ICMR 2008.

[2] H. Bay, T. Tuytelaars, and L. Van Gool, “SURF: Speeded up robust features”, in ECCV 2006.

[3] Sphinx3, http://sourceforge.net/projects /cmusphinx /files/

[4] Seung Eun Lee et. al., “Accelerating mobile augmented reality on a handheld platform,” IEEE International Conference on Computer Design (ICCD), pp.419-426, 4-7 Oct. 2009.

[5] R. Iyer et. al., “CogniServe: Heterogeneous Server Architecture for Large-Scale Recognition,” IEEE Micro, 2011.

이 승 은

1998년 02월 KAIST 전기 및 전자공학과 학사.
2000년 02월 KAIST 전기 및 전자공학과 석사.
2008년 12월 University of California, Irvine, Dept. of Electrical and Computer Engineering, 공학박사.
2010년 09월~현재 서울과학기술대학교, 전자정보공학과.
2009년 03월~2010년 08월 Intel Labs, SoC Architecture Lab.
2008년 07월~2008년 12월 Intel Labs, Oregon Microarchitecture Lab.
2007년 06월~2007년 09월 Broadcom, Wireless connectivity Group.
2006년 06월~2006년09월 Morpho Technology, Soc Design Group.
2000년 02월~2009년 02월 전자부품연구원, SoC 연구센터.
〈관심분야〉 컴퓨터 구조, 다중시스템반도체 (MPSoC), 네트워크 온칩 (NoC)