이동통신 기기용 온/오프라인 통합형 임베디드 필기체 문자 인식 엔진

OSS 지원사업 워크숍

2010. 5 .19

중앙대학교 컴퓨터공학부

휴먼인터페이스 연구실

순 서

- 서론
  - 배경, 개요
- 개발 내용
  - 단계별 개발 내용
  - 1차년도 개발 내용
  - 2차년도 개발 계획 및 방법
- 활동 내용
  - 커뮤니티 운영/관리, 기술 정보 활동
- 요약

배 경

- 모바일 기기의 소형 경량화 및 대중화
- 다양한 서비스에 대한 정보 입력이 필수적
- 숫자키 패드에 의한 정보 입력의 불편함
- 카메라, 터치패드 등의 다양한 입력기기 탑재
- 다양한 환경에 대응하는 통합 문자 인식기 필요

- 실용적인 한글 문자 인식 공개 S/W 없음
- 한글 문자 인식을 위한 데이터의 절대 부족

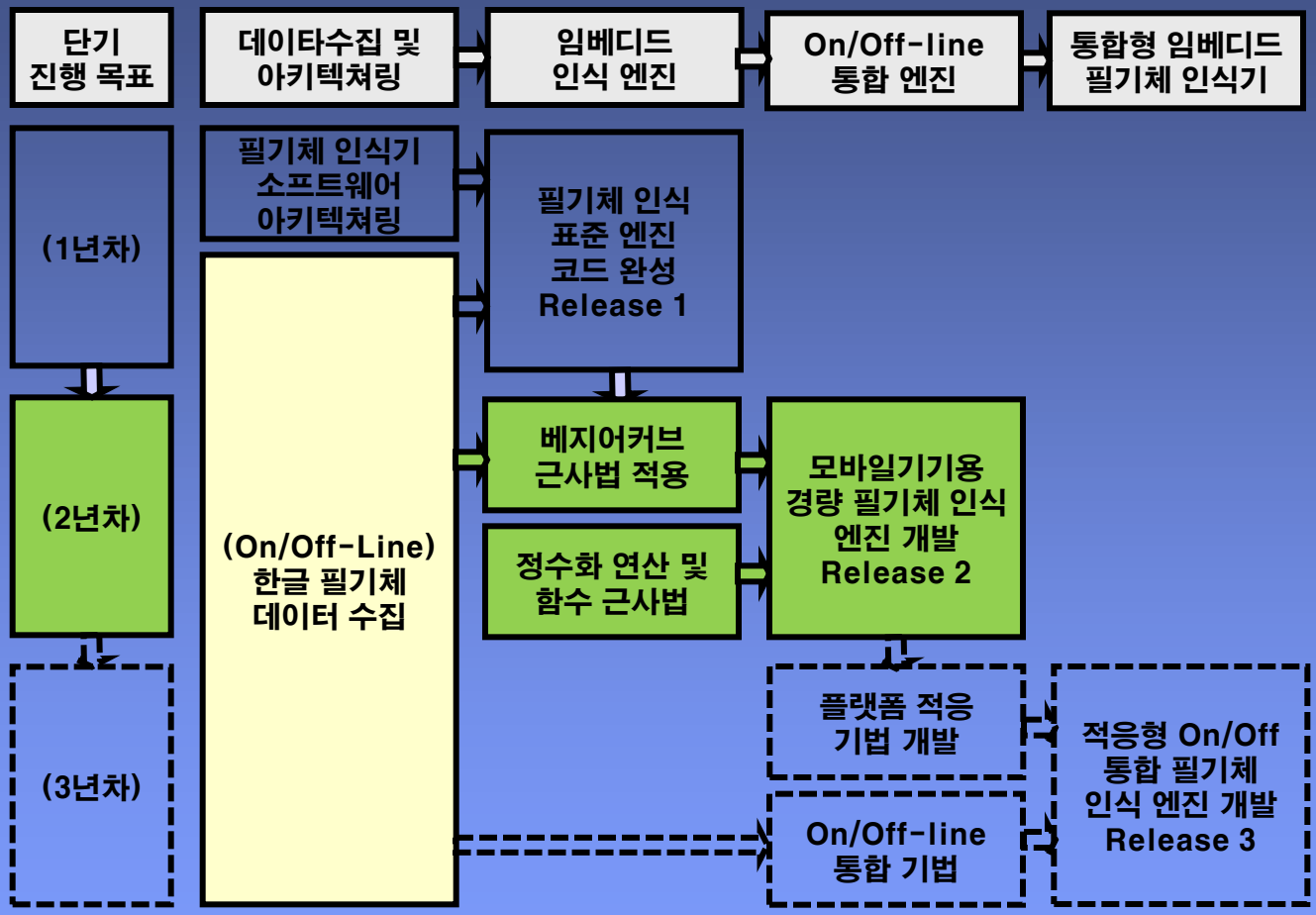
개 요

- **모바일 기기용** 한글 문자 인식 S/W 개발
- 개발용 필기체 **한글 데이터 확보**

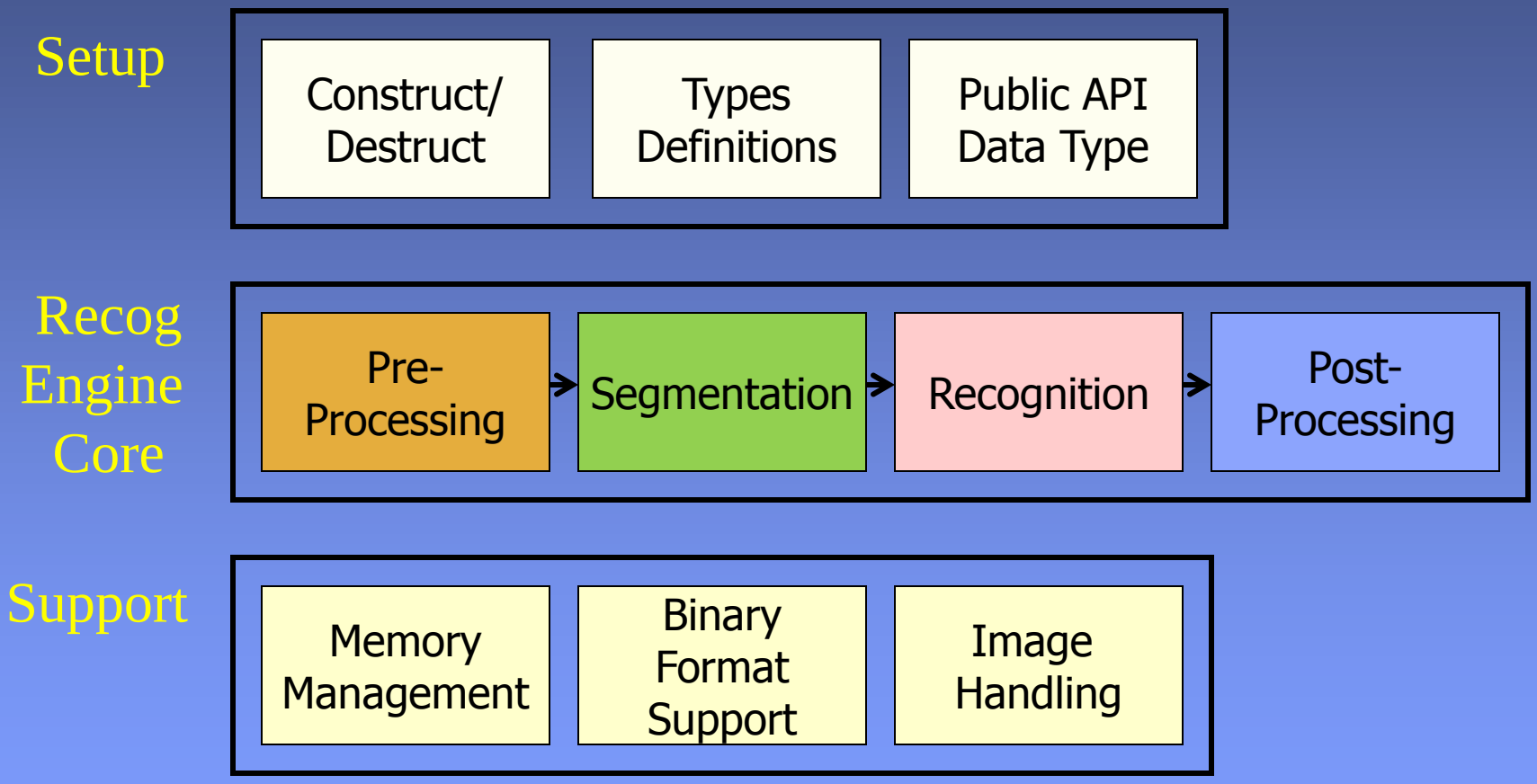
- **통합**형 인식 엔진
  - 온라인/오프라인의 통합 인식
  - 인쇄체/필기체의 통합 인식
  - 다중 언어/문자의 통합 인식

- **임베디드** 모듈화
  - 모바일 기기에 대응하도록 소형 경량화
  - 다양한 플랫폼에 대응하여 이식성 강화
  - 플랫폼 리소스변화에 대응하는 적응성 강화

단계별 개발 내용

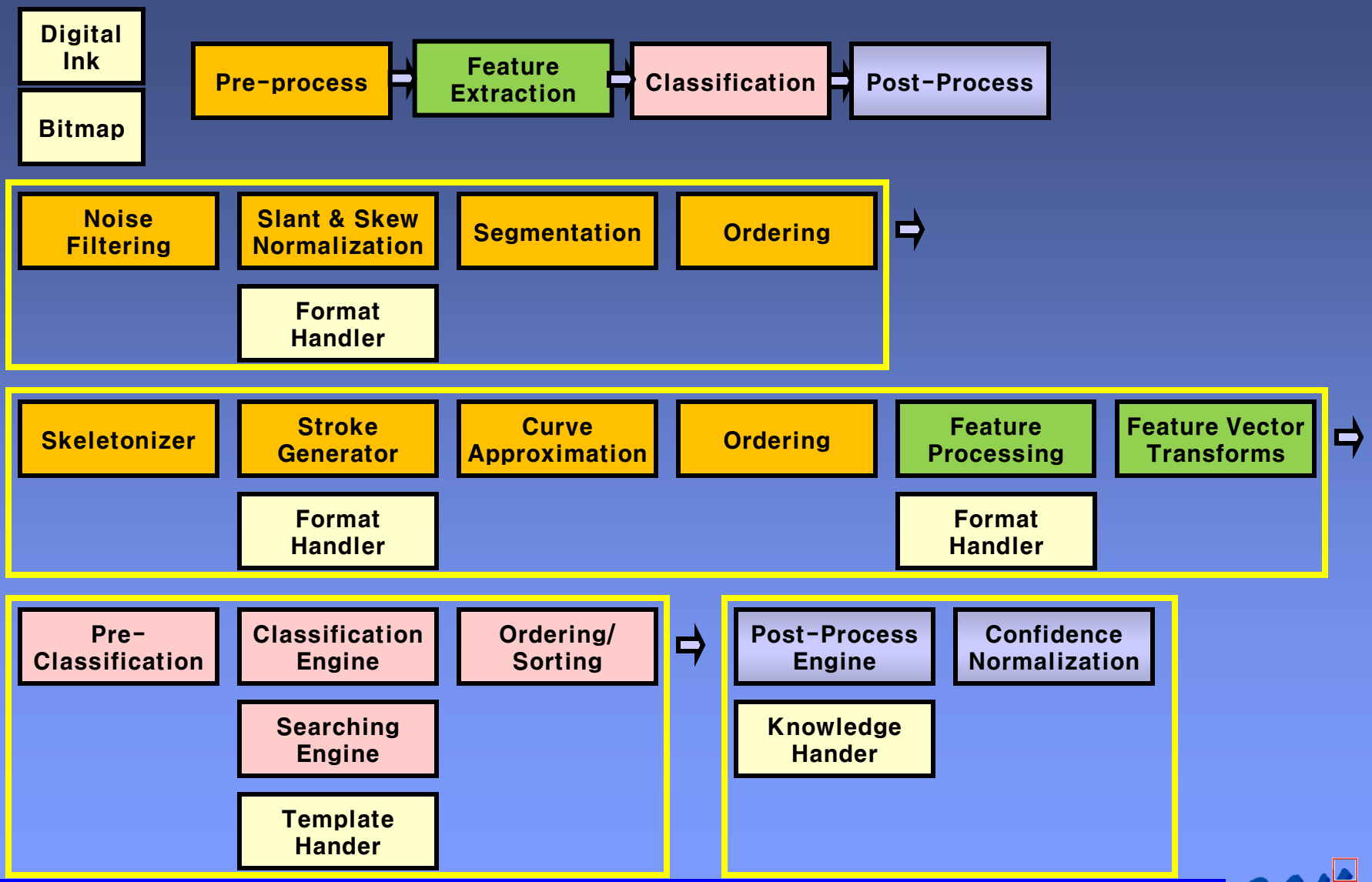

단기
진행 목표
데이타수집 및
아키텍쳐링
임베디드
인식 엔진
On/Off-line
통합 엔진
통합형 임베디드
필기체 인식기
(1년차)
필기체 인식기
소프트웨어
아키텍쳐링
필기체 인식
표준 엔진
코드 완성
Release 1
(2년차)
베지어커브
근사법 적용
모바일기기용
경량 필기체 인식
엔진 개발
Release 2
(On/Off-Line)
한글 필기체
데이터 수집
정수화 연산 및
함수 근사법
(3년차)
플랫폼 적응
기법 개발
적응형 On/Off
통합 필기체
인식 엔진 개발
Release 3
On/Off-line
통합 기법

S/W 구조 - 다중층 구조


Setup
Construct/ Destruct
Types Definitions
Public API Data Type
Recog Engine Core
Pre-Processing
Segmentation
Recognition
Post-Processing
Support
Memory Management
Binary Format Support
Image Handling

S/W 구조 - 모듈 설계

1차년도 개발 내용 2



필기체 문자 데이터 수집

1차년도 수집 분량-1

- 전체 수집 데이터

분류명	영어	숫자	특수문자	한글	총계
온라인	212,355	32,206	6,222	188,459	439,242
오프라인	37,981	7,675	16,029	1,400,042	1,461,727

- 온라인 데이터 수집

구분	영어	숫자	특수문자	한글	총계
Kaist	43,529	3,114	282	90,009	136,934
연세대	0	788	4,133	98,450	103,371
UJI	1,966	397	0	0	2,363
Unipen	166,860	27,907	1807	0	196,574

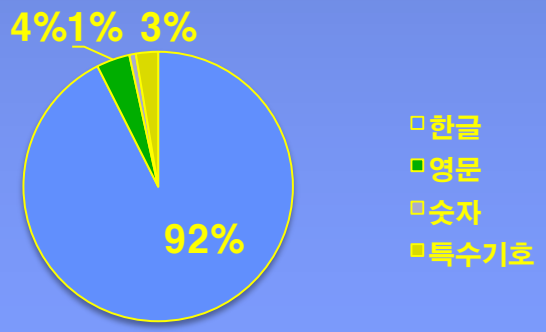
필기체 문자 데이터 수집

1차년도 수집 분량-2

- 오프라인 데이터 수집

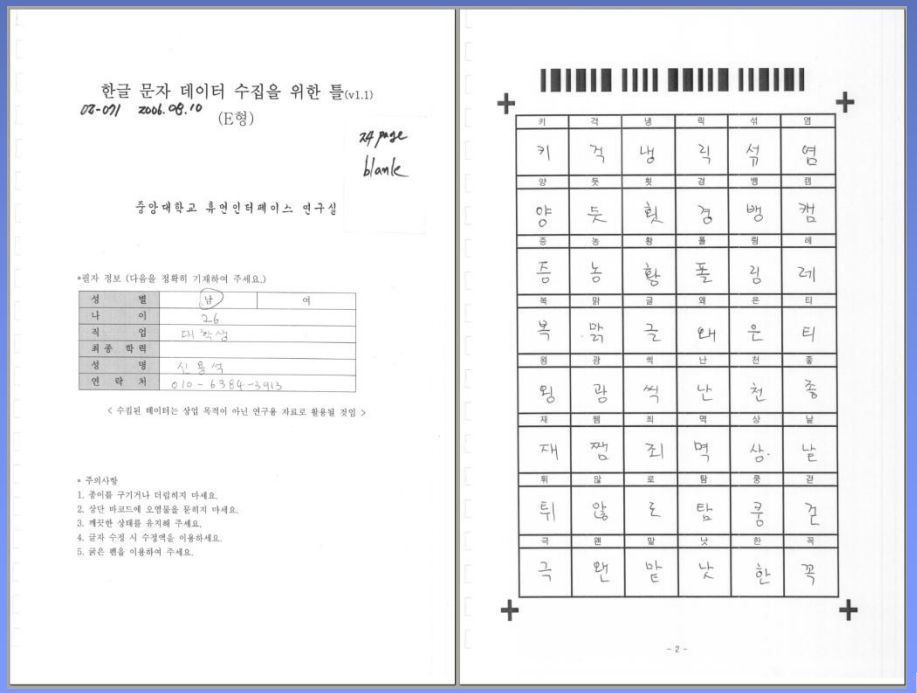
세트번호	부수	수집 문자수	정제 문자수	제거율
SET 1	132	310,200	220,475	28.92%
SET 2	131	333,002	274,494	17.56%
SET 3	77	195,734	192,211	1.79%
SET 4	65	165,230	157,227	4.84%
SET 5	96	244,032	198,948	18.47%
SET 6	71	180,482	179,634	0.46%
SET 7	100	254,200	238,738	6.08%

문자 종류별 비율



오프라인 문자 수집-수집틀

- 한글 2350자, 영문자(대소), 숫자, 특수 문자
- 문자별 품질의 균질성 위해 무작위 배열
- 6가지 서로 다른 문자 배열 순서
- 자료제공자의 기초 자료 포함
- 자동화 인식표 삽입



한글 문자 데이터 수집을 위한 틀(v1.1)

02-071 2006.08.10

(E형)

24 page blank

중앙대학교 휴먼인터페이스 연구실

*필자 정보 (다음을 정확히 기재하여 주세요)

성 별	남	여
나 이	26	
직 업	대학생	
최종 학력		
성 명	신 용 기	
연 락 처	010 - 6386 - 3413	

< 수집된 데이터는 상업 목적이 아닌 연구용 자료로 활용될 것임 >

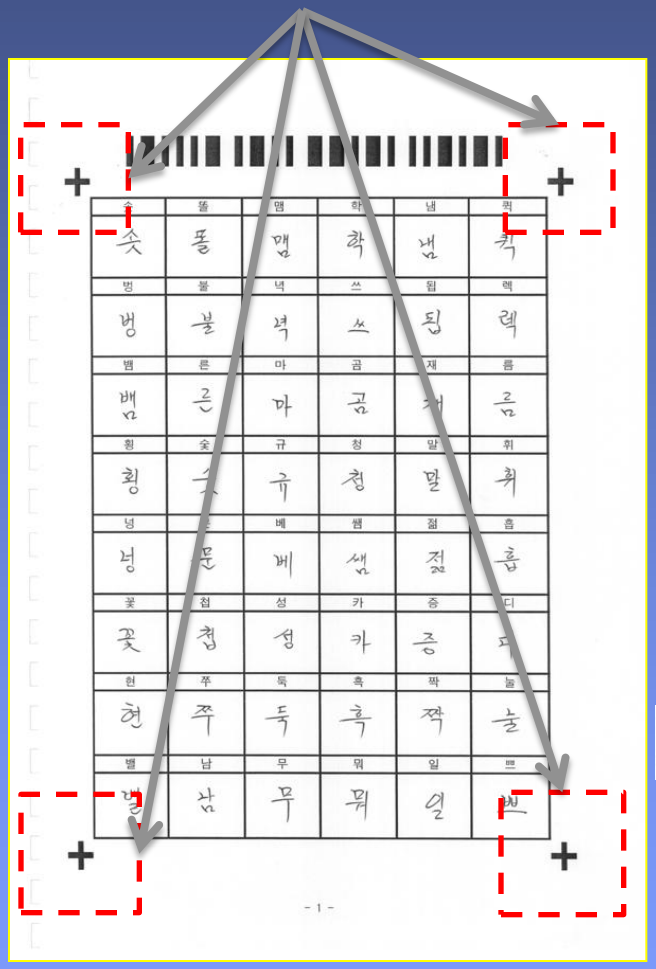
* 주의사항
1. 종이를 구기거나 더럽히지 마세요.
2. 상단 바코드에 오염물을 묻히지 마세요.
3. 깨끗한 상태를 유지해 주세요.
4. 글자 수정 시 수정액을 이용하세요.
5. 굵은 펜을 이용하여 주세요.

키	걱	냉	락	섶	염
양	둣	훳	경	뱅	껌
즘	농	함	폴	링	레
복	맑	글	왜	운	티
뭥	광	썩	난	천	종
재	짬	죄	멱	삼	낱
튀	않	로	탐	궁	걷
극	랜	맡	낫	한	꼭

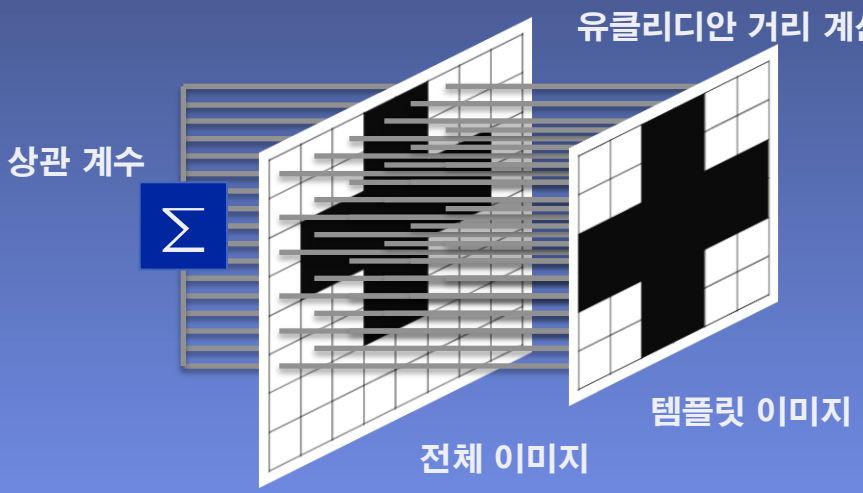
- 2 -

오프라인 문자 수집 - 문자 분리

십자 마커 탐색 영역

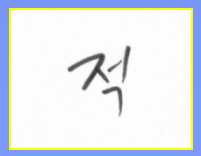


1. 십자 마커 영역 탐색 영역 계산
2. 상관 계수 계산을 통한 십자 마커 탐색

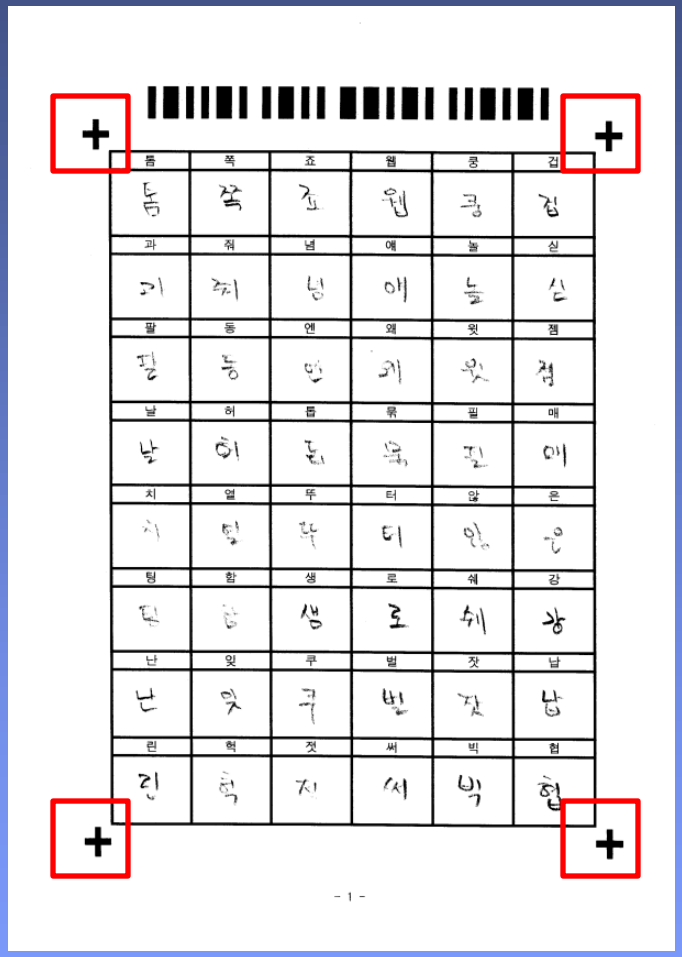


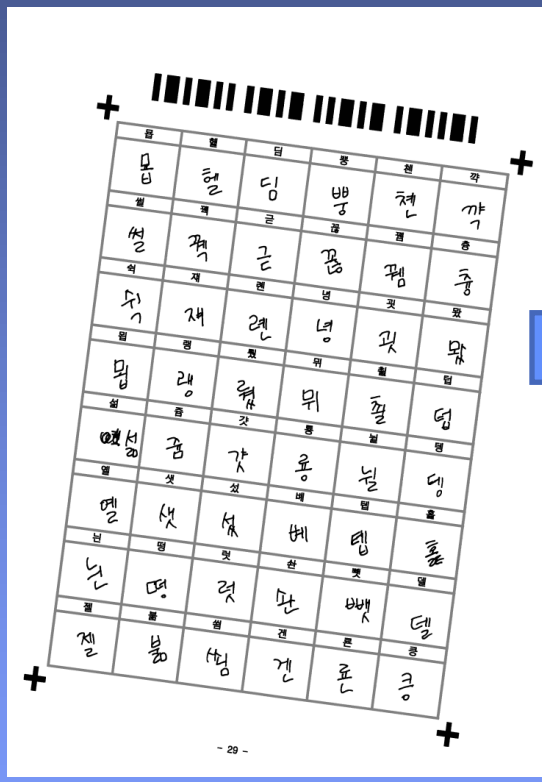
3. 바코드 판독

4. 문자 추출

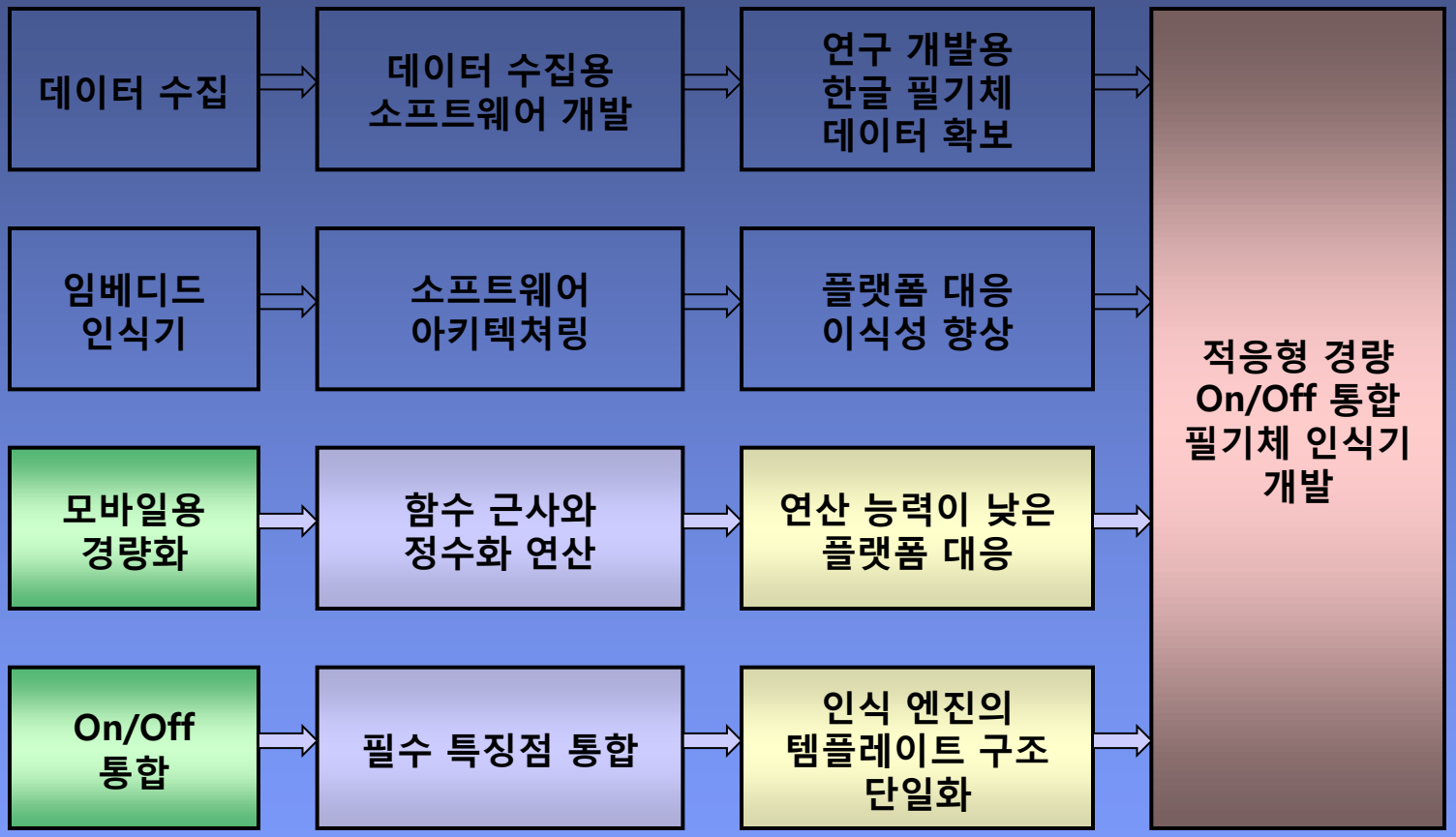


오프라인 문자 수집 - 오차 보정

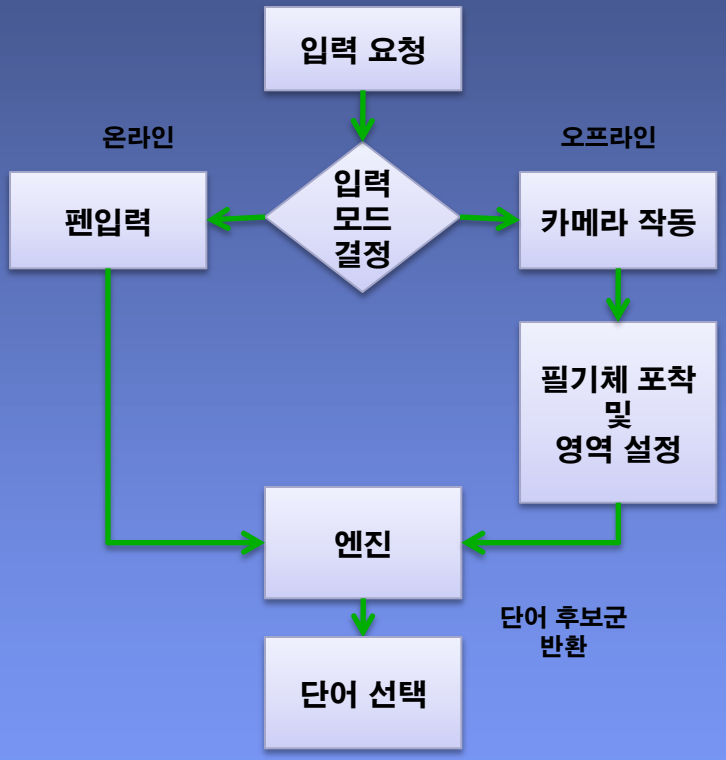




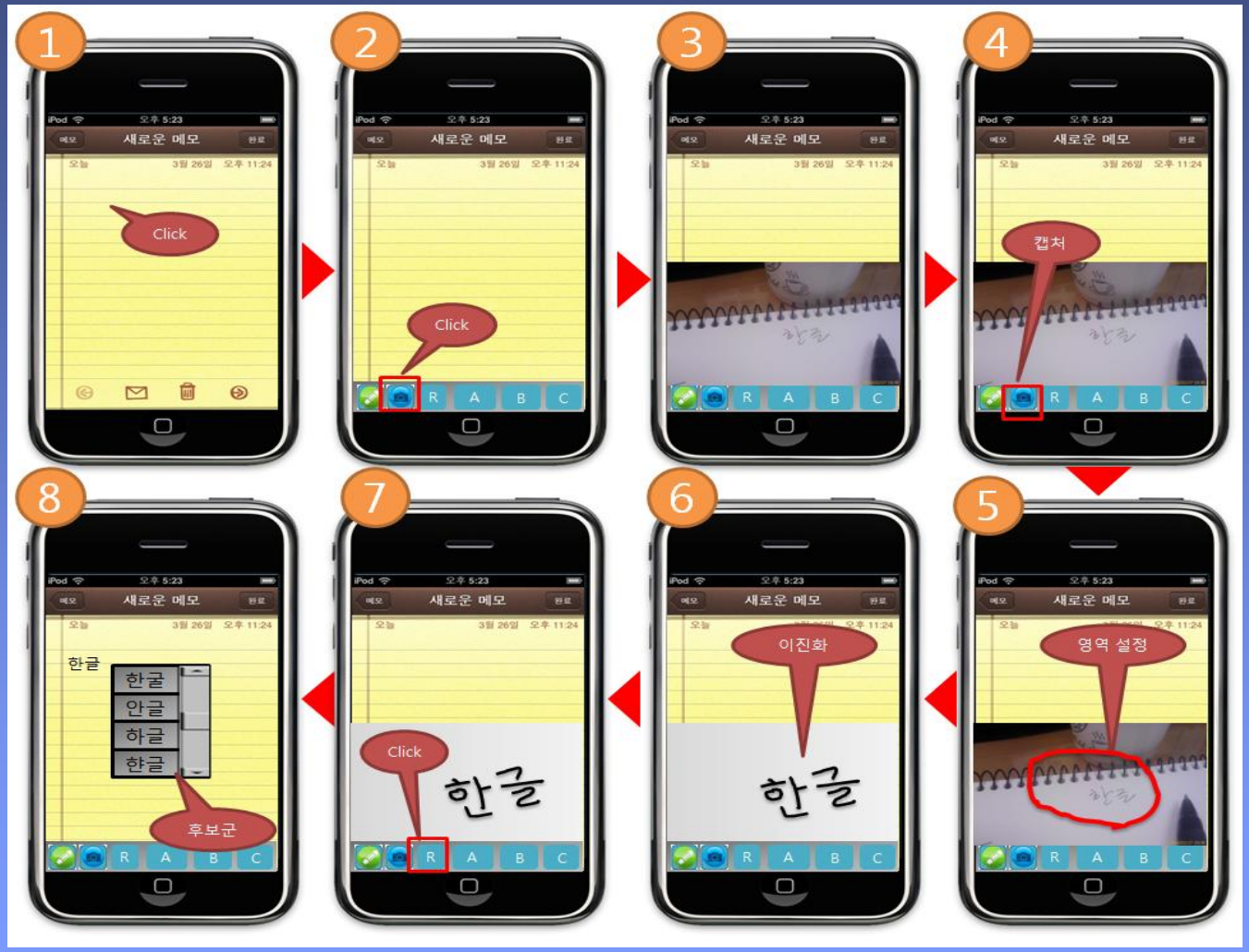
내용과 목표 - 2차년도


데이터 수집
데이터 수집용 소프트웨어 개발
연구 개발용 한글 필기체 데이터 확보
임베디드 인식기
소프트웨어 아키텍쳐링
플랫폼 대응 이식성 향상
모바일용 경량화
함수 근사와 정수화 연산
연산 능력이 낮은 플랫폼 대응
On/Off 통합
필수 특징점 통합
인식 엔진의 템플레이트 구조 단일화
적응형 경량 On/Off 통합 필기체 인식기 개발

사용자 인터페이스 – 통합 인터페이스



사용자 인터페이스 – 오프라인 모드



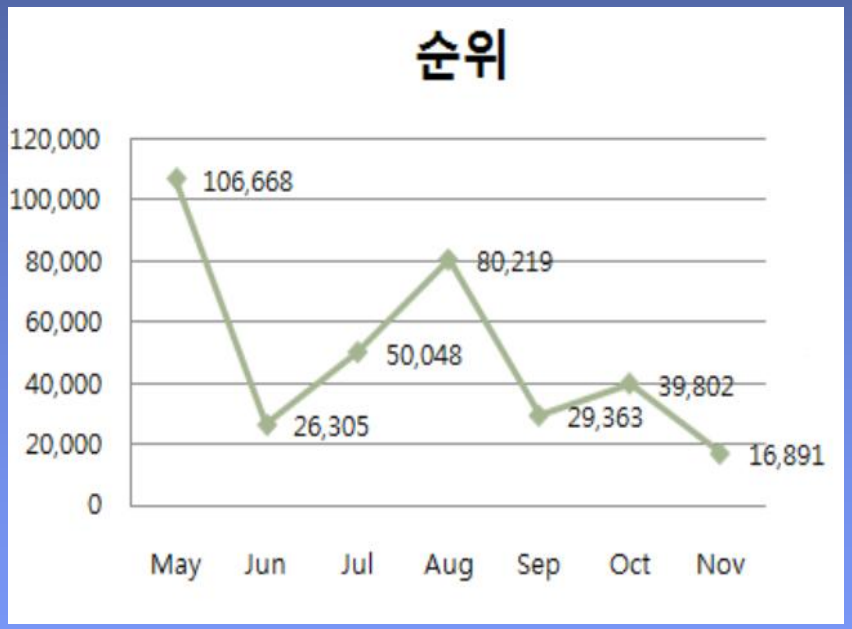
프로그래밍 환경 구축

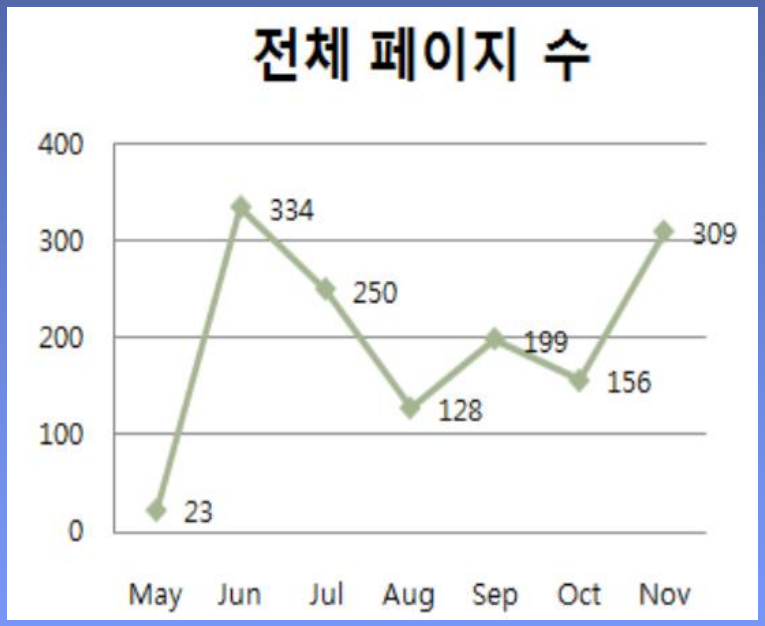
- 개발 환경 조성
  - 개발 디렉터리 구조 설계
  - 소스 컨트롤, Makefile등 빌딩 관련 구조 개발
  - 하위 모듈별 개발 담당자 배정 및 확정

- 임베디드 프로그래밍
  - 이식성향상을 위한 변수형, 선언자 를 규정
  - 함수의 리턴형 통일 및 에러 코드 발생 규칙 설정
  - 코드의 통일성 유지를 위한 코딩 규칙 제정

커뮤니티 운영

- https://sourceforge.net/projects/ucr-project/
- http://kldp.net/projects/ucr/wiki/





기술 정보 활동

- 저작권 관리
  - 저작권 위원회 저작권 관련 전문가 자문 (3회)
  - 공개 소프트웨어 저작권 관련 학술대회 참여
- 과제 홍보 및 협력
  - 디오텍㈜와 문자인식 관련 기술 멘토 및 협력
  - 국제 학술 논문
  - CJKPR 10 논문 발표 (2010년 11월 4-6일, 일본 후쿠오카)
- 교육 기회 제공
  - 패턴인식, 소프트웨어공학 관련 프로젝트
  - 컴퓨터 공학과 졸업 프로젝트로 참여 확대

개발 계획

과제내용	추 진 일 정												참여인력 (M/M)
	1	2	3	4	5	6	7	8	9	10	11	12	
베지어 곡선 특징점 추출													0.3
연산 효율 개선 방안 연구													0.3
사용자 인터페이스 개선													0.3
인쇄체 인식 확장 모듈													0.3
한글 필기체 데이터 수집													0.2
주요 Milestone 완성점에서의 수행결과	● 온/오프라인 통합 표준 특징점 추출 알고리즘 개발 보고서						● 연산 효율 개선 따른 성능 평가 보고서 ● 온/오프라인 통합 필기체 인식 엔진 코드 (release 2)						1.4

요 약

- 모바일용 통합형 필기체 문자 인식 엔진 개발
  - 모바일 기기용 경량 필기체 인식 엔진 개발
  - 모바일 기기용 사용자 인터페이스 개발
- 한글 필기체 데이터의 지속적 수집
  - 오프라인: 200세트 수집, 1000세트 수집 완료(200만 자)
  - 온라인: 기존의 공개 데이터 수집, 규격 통합 재생산 (50만 자)
- 커뮤니티 운영 및 기술 정보 활동
  - 개발 소스 코드, 관련 문서 배포 → 커뮤니티 활성화 기대
  - 관련 기술의 홍보, 저작권 관련 기술 내용 확보