동영상에서 시간 영역 정보를 이용한 자막 검출 알고리듬

論 文
53D-8-9

Caption Detection Algorithm Using Temporal Information in Video

權 哲 鉉* · 申 靑 鎬** · 金 秀 妍*** · 朴 相 暿§
(Chul-Hyun Kwon · Chung-Ho Shin · Su-Yeon Kim · Sang-Hui Park)

Abstract - A noble caption text detection and recognition algorithm using the temporal nature of video is proposed in this paper. A text registration technique is used to locate the temporal and spatial positions of captions in video from the accumulated frame difference information. Experimental results show that the proposed method is effective and robust. Also, a high processing speed is achieved since no time consuming operation is included.

Key Words : Text Detection, Video Indexing

1. 서 론

최근에 디지털 영상이 빠르게 증가하면서 디지털 영상의 자동화 컨텐츠를 기반으로 하는 인덱싱이 주목을 받고 있다. 영상에 첨가된 자막 문자는 비디오 인덱싱과 검색에 유용한 정보를 제공한다. 영상에 첨가된 자막 문자는 영상의 내용을 어느 정도까지 직접적으로 묘사한다. 예를 들어, 뉴스 영상에서 자막은 보통 위치나 관련된 사람의 이름, 그리고 뉴스에 관련된 특수한 주제들을 포함한다.

영상에서 문자를 추출하는 방법들이 많이 제안되었지만, 대부분은 시간상의 정보를 고려하지 않고 각각의 비디오 프레임을 독립적으로 다루었다[1]-[4]. 단지 멀티프레임 통합을 사용하는 문자 향상 부분에서만 시간상의 분석이 적용되었다[2]. Tang은 시간상의 정보를 이용해 자막 전이가 발생하는 프레임의 위치를 찾았다[5]. 그러나 자막 전이를 찾는 기본적인 기준이 프레임 간의 차이이므로, 이러한 방법들은 움직이는 물체의 속도나 카메라 움직임이 어느 범위 이상으로 변하면 자막 전이를 찾지 못한다. 또한 자막 전이를 찾을 때 발생하는 잘못된 검출(false alarm)과 잘못된 미검출(false rejection)이 자막이 사라지는 프레임과 자막이 나타나는 프레임을 찾기 어렵게 만든다.

이 논문에서는 자막 문자의 장시간의 동작에 초점을 두었다. 먼저, 문자 출현 맵(text appearance map)을 구성하기 위해 문자 등록 기술을 소개한다. 그리고 전통적인 방법들로 문자 출현 맵에서 문자의 시공간상의 위치들을 추출한다. 그런 후, 문자 향상을 위해 멀티프레임 통합이 사용되고, 이분화 과정이 적용된다.

* 正 會 員 : 延世大 電氣·電子工學科 博士課程
** 正 會 員 : 延世大 電氣·電子工學科 博士課程
*** 學生會員 : 延世大 電氣·電子工學科 碩士課程
§ 正 會 員 : 延世大 電氣·電子工學科 敎授 · 工博
接受日字 : 2004年 3月 22日
最終完了 : 2004年 5月 30日

2. 본 론

2.1 자막 검색

2.1.1 문자 등록 기술

제안된 방법의 기본적인 생각은 변화의 검색이다. 그러나 우리의 자막 검색의 판단 기준은 연속적인 두 프레임의 차이가 아니라, 일련의 영상으로부터 최근의 문자 등록 정보를 유지하여 구성한다. 전 프레임과 큰 차이가 있고 오랜 기간 동안 지속성을 가지는 픽셀은 그림1에서 보듯이, 그 시간 동안에 자막 문자로 추정된다.

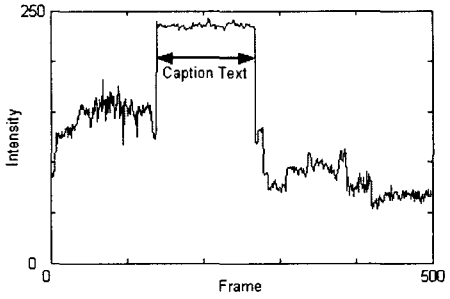


그림 1 픽셀 밝기의 시간에 따른 변화

Fig. 1 The temporal intensity profiles of a pixel

이 정보는 움직이는 물체와 카메라 움직임이 있는 경우에 더 믿을만하다. 우리는 이를 위해 하나의 메모리를 알고리듬에 추가했다. 메모리의 값들은 자막 문자로 추정된 픽셀들이 몇 번의 프레임에서 연속적으로 나타나는 지를 가리킨다. 메모리는 표1에 나타난 기준에 따라 각각의 프레임에서 갱신된다. MEM_k는 k번째 프레임에서의 메모리, |FD|는 프레임 차이의 절대값, L은 미리 정의된 수, TH_{FD}는 변화 검색을 위한 경계값이고 중요도 테스트(Significance test) 기술[6]로 구한다. 프레임 차이 영상에서 해당하는 픽셀에서의 값이 TH_{FD}를 초과하면 그 픽셀은 주요한 변화(Significant Cha-

ngc) 로 분류된다. MEM_0의 모든 값들은 0으로 초기화된다. 우리는 [5]에서 제안된 방법으로 각각의 프레임들이 셧 경계인지 아닌지를 조사했다. 셧 경계는 프레임 차이에서 상대적으로 큰 값을 가지지만 이러한 변화의 대부분은 자막과는 무관하다. 자막은 장면 전환(3-4, 6의 경우)동안 항상 안정적이기 때문에 셧 경계에서는 단지 시간상의 일관성만을 고려한다.

표 1 자막 테스트 판단 기준

Table 1 Caption Decision Criteria

Case	장면 전환	프레임 차이	이전 메모리	현재 메모리	비고
1	No	$\|FD\| \le TH_{FD}$	MEM=0	MEM=0	비자막영역
2	No	$\|FD\| > TH_{FD}$	MEM≤L	MEM=1	후보영역
3	Yes	$\|FD\| > TH_{FD}$	MEM≤L	MEM=0	비자막영역
4	Any	$\|FD\| \le TH_{FD}$	MEM>0	MEM+=1	후보영역
5	No	$\|FD\| > TH_{FD}$	MEM>L	MEM=1	자막등록
6	Yes	$\|FD\| > TH_{FD}$	MEM>L	MEM=0	자막등록

1의 경우는 메모리의 값이 0으로 초기화 된 이후 프레임 차이에서 중요한 변화가 일어나지 않은 경우를 나타낸다. 현재 해당 블록이 후보 자막 텍스트 상태가 아니므로 다음 메모리 값에 변화가 생기지 않는다. 2의 경우는 프레임 차이에서 중요한 변화가 생겼지만 일정 기간동안 지속되어야할 시간적 연속성을 만족시키지 못하는 경우이다. 자막 텍스트로 등록되기 위해서는 후보 자막 텍스트로 설정한 후 최소한 L 프레임이상 픽셀 값의 변화가 없어야 한다. 이 경우 메모리 값은 1로 설정한다. 0이 아니라 1로 설정하는 이유는 현재 프레임 차이에서의 중요한 변화가 자막 텍스트의 출현에 의해서 생길 수도 있기 때문이다. 3의 경우는 장면전환에 대한 조건을 제외하면 2의 경우와 유사하다. 이 경우도 2의 경우와 마찬가지로 프레임 차이에 중요한 변화가 있지만 시간적 연속성을 만족시키지 못하고 있다. 그러나 2와는 다르게 프레임 차이가 자막 텍스트가 아닌 장면 전환으로 생긴 것이므로 메모리 값은 0으로 초기화 한다. 4의 경우는 후보 자막 텍스트 상태였던 블록이 시간적 연속성을 만족시키는 경우이다. 즉, 장면 전환 여부와는 상관없이 후보 상태였던 블록의 밝기 값이 이전 프레임에 비해 변화하지 않았기 때문에 1만큼 증가시킨다. 5와 6의 경우에는 후보 자막 텍스트 상태인 블록이 L프레임 이상 지속된 후 프레임 차이에서 중요한 변화가 발생했으므로 자막을 나타내는 블록으로 등록된다. 5의 경우는 자막 블록이 잘못 등록되어진 경우 메모리 값을 0으로 설정하면 바로 이어서 나타나는 자막 블록을 등록시키지 못하는 경우가 생기므로 메모리 값을 1로 설정된다. 반면에 6의 경우는 프레임 차이에서의 중요한 변화가 장면 전환에 기인하기 때문에 메모리 값은 0으로 설정된다.

모든 픽셀들에 대해 과정이 끝난 후에는 계산량을 줄이기 위해서 문자로 등록된 픽셀들의 수를 센다. 등록된 픽셀의 수가 경계값, T_r을 초과하면, MEM에 등록된 픽셀들의 값들은 다음과 같이 문자 출현 맵(TAMs)에서 상응하는 픽셀들에 복사된다.

$$TAM_{k-MEM_k(i,j)}(i,j) = MEM_k(i,j) \qquad (1)$$

k는 현재의 프레임 넘버이다. TAM_K는 k번째 프레임에만 나타나는 등록된 자막 픽셀들에 대한 정보를 가진다. 메모리에서 복사된 TAMs의 픽셀 값들은 상응하는 자막 픽셀들의 기간을 나타낸다. 등록되지 않은 픽셀들의 값들은 0이다. 그림 2는 문자 등록 결과들을 보여준다. 그림 2(d)는 주요한 변화로 분류된 픽셀들을 흰색으로 둔 주요한 변화 맵을 보여준다. 그림 2(e)에서 보여준 것처럼 자막 영역은 TAM에서 나타난다.

(a)이전영상 (b)다음영상 (c)프레임 차이 (d)주요한 변화 맵 (e)(b)에서의 TAM (f)자막 텍스트 검출 결과

그림 2 자막 텍스트 검출 전과정

Fig. 2 Overview over the caption detection step

2.1.2 자막 라인 위치 검출

이 단계에서는 각각의 TAM에서 자막의 시공간상의 위치를 찾는다. 연속적으로 나타나는 자막들은 한 프레임씩 검출하게 되면, 그림 3에서 보듯이 글자가 잘려서 검출이 불가능하거나 검출이 가능하다고 해도 글자 단위로 검출이 되기 때문에 색인이나 검색을 위하여 하나의 단어나 문장으로 다시 조합해야만 한다. 매 프레임마다 자막 텍스트를 인식한 후 추적 과정을 거쳐 중복된 자막 텍스트를 제거하는 기존의 연구들은 이러한 문제를 해결할 수 없었다. 그림 4에서 보듯이 점진적으로 나타나는 자막 텍스트의 경우에는 자막이 여러 프레임에 걸쳐서 나타나기 때문에 한 프레임씩 검출할 경우 잘못 검출될 가능성이 매우 높다. 이러한 문제를 해결하기 위해 본 연구에서는 다음과 같은 과정을 거친다. 연속적(점진적)으로 나타나는 자막 텍스트의 경우 TAMs에서 연속적으로 등록된 픽셀의 수가 큰 값을 갖는다. $Nr(TAM_K)$을 TAM_K에 등록된 총 픽셀의 수와 같다고 두었다. 그리고 그 값들을 좀 더 정확하게 위치시키기 위해서, $Nr(TAM_K)$과 $Nr(TAM_{K+1})$이 모두 정의된 어떤 문턱값 Tr보다 크면, 식 (2)을 이용해 TAM_K에 등록된 픽셀들의 정보를 다음과 같이 TAM_{K+1}로 이동시킨다.

$$TAM_{k+1}(i,j) \mathrel{+}= TAM_k(i,j) - 1 \text{ , if } TAM_k(i,j) > 0 \qquad (2)$$

k는 프레임 넘버이고, i, j는 프레임에서의 좌표를 의미한다.

(a) 545th frame (b) 553th frame (c) 559th frame

그림 3 연속적으로 나타나는 자막 텍스트의 예
Fig. 3 Sequentially appearing caption examples

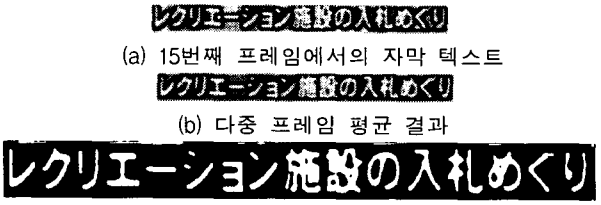
(a) 25th frame (b) 29th frame (c) 33th frame

그림 4 점진적으로 나타나는 자막 텍스트의 예
Fig. 4 Gradually appearing caption examples

그리고 대략적인 프로젝션 프로파일(projection profile) 방법을 사용하여 대략적으로 후보 자막 영역을 찾았다. 이렇게 찾은 후보 자막 영역은 하나의 자막 라인만을 가지고 있지는 않다. 또한 카메라의 이동이나 물체의 움직임에 의해 잘못 검출되는 경우도 포함하고 있기 때문에 잘못 검출된 후보 자막 영역을 제거하고, 여러 개의 자막 라인으로 이루어진 후보 자막 영역을 하나의 자막 라인으로 분리하기 위해서 [3]에서의 지역 분해 방법을 사용한다. 그리고 이렇게 TAM을 이용하여 찾은 자막 텍스트 영역에서 등록된 픽셀 값의 히스토그램을 이용하여 자막 텍스트가 사라지는 프레임을 찾고, 완전히 제거되지 않은 잘못 찾아진 자막 텍스트 영역도 제거한다. TAM에 등록되어 있는 픽셀의 값은 해당 픽셀이 얼마나 오랫동안 자막 텍스트로 유지되는 지를 나타내는 것으로 검출된 자막 텍스트 영역이 진정한 자막 텍스트라면 그 영역에 등록되어 있는 대부분의 픽셀들의 값은 같을 것이라고 예상할 수 있다. 이는 자막 텍스트에 해당하는 픽셀들은 동시에 나타났다가 동시에 사라지기 때문이다. 이러한 정보를 이용하면 후보 자막 텍스트 영역 내부에 자막 텍스트가 존재하는 지 여부와 존재한다면 그 자막 텍스트가 언제 사라지는 가를 알 수 있다. 따라서 후보 자막 텍스트 영역에서 등록된 픽셀에 대한 0의 값을 제외한 TAM의 히스토그램 H[i], i=1,2,...,N 을 계산한다. 그리고 H(n)를 히스토그램의 최대값이라고 두고, 그 값이 차지하는 부분이 경계값보다 크면, 문자 박스에 있는 자막 문자는 n개의 프레임 후에 사라진다고 추정되고, 그렇지 않으면 후보 자막 텍스트는 잘못된 검출로 간주하고 제거된다. 그림 2(f)는 자막 라인 위치 검출의 결과이다.

2.2 자막 향상과 인식

비디오 영상에서 해상도가 낮은 문자와 복잡한 배경은 문자 인식을 상당히 어렵게 만든다. 후자의 문제를 해결하기 위한 방법으로는 다중 프레임의 평균값이나 최소(혹은 최대) 픽셀 검색이 주로 사용된다. 이러한 방법들을 실행할 때, 중요한 것은 같은 문자열을 가진 자막 라인의 최대값을 등록하는 것이다. 제안된 방법은 전 섹션에서와 같이 할 수 있다. 그림 5(b)는 다중 프레임 평균을 통한 자막 향상의 예다. 전자의 문제를 해결하기 위해, 서브 픽셀 보간 기술[2]을 사용해 해상도가 낮은 문자들을 보간하였다. 그 다음에, 보간된 영상은 등록된 자막 픽셀들의 평균에서 얻은 고정된 한계값으로 이진화된다. 그림 5(c)와 (d)는 문자 이진화와 인식 결과의 예들이다.



(a) 15번째 프레임에서의 자막 텍스트

(b) 다중 프레임 평균 결과

(c) 보간 후에 이진화된 텍스트

しクリエーション鎭設の入札めくリ

(d) 문자 인식 결과

그림 5 자막 강화와 이진화의 예
Fig. 5 Caption enhancement and binarization examples

3. 실험과 결과

실험 데이터는 한국의 KBS 뉴스 영상과 일본의 NHK 뉴스 영상을 사용하였다. 5시간의 뉴스 영상들이 해상도가 320*240인 MPEG-1으로 코딩되었다. 이 데이터들은 1821개의 자막 라인을 가지고 있다. 연속적인 모든 프레임들을 사용하는 대신, 점진적으로 나타나는 자막을 좀 더 정확하게 찾기 위해 고정된 간격(=2)을 두고 프레임들을 처리했다. 실험에서는 L=70(즉 2.3초), Tr=400으로 두었다. 이 방법의 실험 결과는 다음과 같이 정의된 리콜(recall)과 정확도(precision)로 보여진다.

$$Recall = \frac{\text{정확하게 찾아낸 검출수}}{\text{정확하게 찾아낸 검출수} + \text{못 찾은 검출 수}} \quad (3)$$

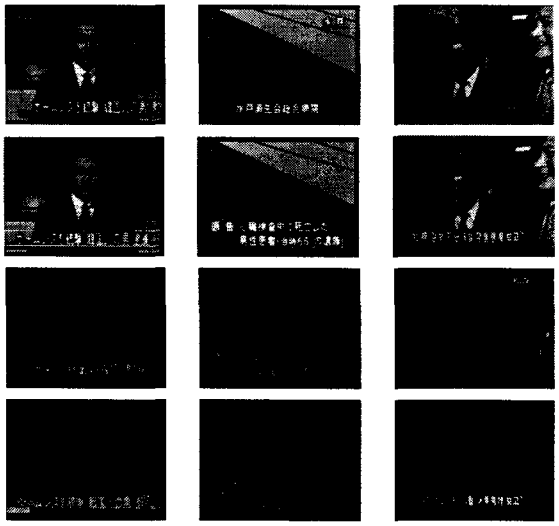
$$\text{Pr}ecision = \frac{\text{정확하게 찾아낸 검출수}}{\text{정확하게 찾아낸 검출수} + \text{잘못 찾아낸 검출수}} \quad (4)$$

3.1 자막 검색 결과

제안된 알고리듬은 영상에 포함된 시간상의 정보를 사용한 Tang의 방법과 비교하였다. 자막 검색 성과에 대한 기준이 명확하지 않기 때문에, 비교를 하기 위해 Tang이 보여준 자막 전이 검색과 자막 라인 위치 추정 실험을 하였다. 자막 전이 검색 실험에서는 792개의 자막 전이가 포함된 942개의 비디오 샷(총 2시간)이 사용되었다. 제안된 방법에서는 하나의 자막 출현 프레임이 검색된다면, 그것에 상응하는 사라지는 프레임 또한 검색된다고 고려하여 TAM과 Frame Difference Map(FDM)을 비교하기 위해 자막 출현 위치에서의 378개의 자막 라인을 포함하는 200개의 프레임을 선택하였다. 그 결과는 표2와 같다. 제안된 접근 방법이 검색율을 향상시키고, 잘못된 검출을 상당히 감소시킨다는 것을 알 수 있다. Tang의 방법은 카메라 움직임에 매우 민감할 뿐만 아니라, 잘못된 검출과 잘못된 미검출이 발생했을 때, 자막 소멸 프레임과 상응하는 자막 출현 프레임을 찾는 방법을 설명하지 않았다. 그러나 제안된 방법에서는 매칭 과정은 필요하지 않다. 표 3은 모든 실험 데이터에 대한 제안된 방법의 평가 결과이다. 98% 이상의 자막 라인이 정확하게 검색되었고, 95%의 정확도를 보였다. 실험 결과는 제안된 방법이 점진적으로 나타났다가 사라지는 자막을 처리할 수 있고, 빠르

게 움직이는 물체와 카메라 움직임에 비교적 강력하다는 것을 보여준다. 또한 제안된 방법은 문자의 구조적인 특징들을 사용하지 않기 때문에, 어떤 타입의 문자, 폰트 스타일, 언어일지라도 검색할 수 있다. 그리고 찾지 못한 자막 라인들은 대부분은 나타날 때의 프레임 차이가 미비하거나 주밍(zooming)과 같은 특수 그래픽 효과로 인한 것이다. 그림 6은 Tang의 방법으로는 검색할 수 없지만 제안된 방법으로는 검색이 되는 예들이다.

인식 실험에서는 792개의 문자를 포함하는 100개의 자막 라인을 수동적으로 선택하였다. 문자 인식은 상업적인 OCR 프로그램인 ARMI PRO로 하였다. 이진 문자에 대한 평균 인식률은 84%였다.



(a) (b) (c)

그림 6 자막 텍스트 검출의 예 : 처음 두 줄은 연속된 프레임을 보여준다. 세 번째 줄은 두 프레임의 차이 영상을 보여준다. 마지막 줄은 각각의 TAM을 보여준다. (a)는 왼쪽에서 오른쪽으로 나타나는 자막, (b)는 급작스럽게 바뀌는 자막, (c)는 점진적으로 나타나는 자막이다

Fig. 6 Our caption detection examples : Adjacent frame pairs are shown in the first two rows. Their difference images are shown in the third row. The corresponding TAMs are shown in the last row. (a) is the caption flying leftward and then disappearing, (b) is an abrupt change of caption and (c) is a caption appearing gradually.

표 2 자막 텍스트 검출 결과 비교

Table 2 Comparison of Caption Detection Results

		전체수	옳게 검출한 수	검출 못한 수	잘못 검출한 수	Recall	Precision
자막 변이 검출	Tang	792	663	129	276	0.84	0.71
	제안한 방법		780	12	44	0.98	0.95
자막 라인 위치	Tang	378	368	10	45	0.97	0.89
	제안한 방법		375	3	2	0.99	0.99

표 3 자막 텍스트 검출 결과

Table 3 Results of Caption Line Detection

	전체수	옳게 검출한 수	검출 못한 수	잘못 검출한 수	Recall	Precision
자막라인 위치	1821	1793	28	97	0.98	0.95

3.2 검출 속도

제안된 방법은 텍스쳐 특징을 추출하거나 텍스트 영역을 찾기 위해 에지 정보를 계산하는 것과 같은 시간을 소비하는 단계가 없다. 또한, 영상에서 모든 프레임에서 문자 검색을 하는 것도 아니고 문자 블록 추적도 필요하지 않다. OCR 과정과 MPEG 디코딩 시간을 제외하면 제안된 방법이 PIII-700에서 하나의 프레임을 처리하는데 걸리는 시간은 0.009가 걸렸다. 하나의 MPEG 프레임을 디코딩하는 데 걸리는 시간은 평균 0.007초이다. 매번 두 프레임당 하나가 처리되기 때문에, 제안된 방법은 실시간보다 약 2.9배 빠르게 MPEG-1 파일로부터 자막 영역을 찾을 수 있다.

4. 결 론

영상에서 이용할 수 있는 시간상의 정보를 통합한 새로운 자막 검색 알고리듬을 제안하였다. 자막의 장기간 시간상의 움직임을 이용하고 프레임 차이의 기록에 의해 갱신된 다이내믹 메모리를 채택한, 문자 등록 기술을 개발하였다. 자막 라인들은 TAM으로부터 쉽게 검색되었다. 실험 결과들은 제안된 방법이 정확하고 강력하다는 것을 보여주었다.

참 고 문 헌

[1] H. Li, D.Doermann, and O. Kia, "Automatic Text Detection and Tracking in Digital Video", IEEE Trans. on Image Processing, Vol. 9, pp. 147-256, Jan. 2000

[2] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for Digital News Archives", in Proc. IEEE Int. Workshop on Content-Based Access of Image and Video Database(CAVID'98), pp. 52-60, 1998.

[3] X. S. Hua, X. R. Chen, L. Wenyin, H. J. Zhang, "Automatic Location of Text in Video Frames", Proceeding of ACM Multimedia 2001 Workshops : MIR2001, pp. 24-27, Ottawa, Canada, October 5, 2001.

[4] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic Caption Localization in Compressed Video", IEEE Trans. on PAMI, Vol. 22, No. 4, pp. 385-392, April, 2000.

[5] X. Tang, X. Gao, J. Liu, and H. J. Zhang, "A Spatial-Temporal Approach for Video Caption Detection and Recognition", IEEE Trans. on Neural Network, Vol. 13, No. 4, pp. 961-971, July 2002.

[6] T. Aach, A. Kaup, and R. Mester, "Statistical model-based change detection in moving video", Signal Processing, Vol. 31, pp. 165-180, Mar. 1993.

저 자 소 개

권 철 현 (權 哲 鉉)

1975년 1월 9일생. 1997년 연세대 전기공학과 졸업. 2000년 동 대학원 전기·전자공학과 졸업(석사). 2000년~현재 동 대학원 전기·전자공학과 박사과정

Tel : 02-2123-2768, Fax : 02-312-7735

E-mail : cherni99@hanmail.net

신 청 호(申 菁 鎬)

1973년 12월 1일생. 1996년 연세대 전기공학과 졸업. 1998년 동 대학원 전기공학과 졸업(석사). 1998년~현재 동 대학원 전기·전자공학과 박사과정

Tel : 02-2123-2768, Fax : 02-312-7735

E-mail : naice@dreamwiz.com

김 수 연 (金 秀 妍)

1980년 1월 28일생. 2002년 경원대 전자공학과 졸업. 2002년~현재 동 대학원 전기·전자공학과 석사과정

Tel : 02-2123-2768, Fax : 02-312-7735

E-mail : ksystory@hotmail.com

박 상 희(朴 相 暿)

1939년 8월 25일생. 1962년 연세대 전기공학과 졸업. 1964년 동 대학원 전기공학과 졸업(석사). 1971년 동 대학원 전기공학과 졸업(박사). 1982년~1984년 Washington Univ. 방문교수. 1998년~1999년 대한전기학회 회장. 1970년~현재 연세대 전기·전자공학과 교수. 2002년~현재 연세대 정보대학원 원장

Tel : 02-2123-2768, Fax : 02-312-7735

E-mail : psh@yonsei.ac.kr